SELECTIVE INFERENCE AND LEARNING MIXED GRAPHICAL MODELS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTATIONAL MATH AND ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jason Dean Lee

May 2015

Jason Dean Lee

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.



(Trevor J. Hastie) Principal Co-Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Jonathan E. Taylor) Principal Co-Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Lester Mackey)

Approved for the Stanford University Committee on Graduate Studies

Abstract

This thesis studies two problems in modern statistics. First, we study selective inference, or inference for hypothesis that are chosen after looking at the data. The motiving application is inference for regression coefficients selected by the lasso. We present the Condition-on-Selection method that allows for valid selective inference, and study its application to the lasso, and several other selection algorithms.

In the second part, we consider the problem of learning the structure of a pairwise graphical model over continuous and discrete variables. We present a new pairwise model for graphical models with both continuous and discrete variables that is amenable to structure learning. In previous work, authors have considered structure learning of Gaussian graphical models and structure learning of discrete models. Our approach is a natural generalization of these two lines of work to the mixed case. The penalization scheme involves a novel symmetric use of the group-lasso norm and follows naturally from a particular parametrization of the model. We provide conditions under which our estimator is model selection consistent in the high-dimensional regime.

Acknowledgements

- I would like to thank my advisors Trevor Hastie and Jonathan Taylor. Trevor has given me the perfect amount of guidance and encouragement through my PhD. I greatly benefited from his numerous statistical and algorithmic insights and input on all my projects. Jonathan has been a great mentor throughout my PhD. He has been generous with sharing his ideas and time. With Yuekai Sun, we have spent countless afternoons trying to understand Jonathan's newest ideas.
- I would like to thank Lester Mackey for serving on my oral exam committee and reading committee.The Stats-ML reading group discussions also introduced me to several new areas of research. I would also like to thank Andrea Montanari and Percy Liang for their spectacular courses and serving on my oral exam committee.
- Emmanuel Candes, John Duchi, and Rob Tibshirani have always been available to provide advice and guidance.
- Yuekai Sun and I have collaborated on many projects. I've benefited from his numerous insights, and countless discussions. I am lucky to have a collaborator like Yuekai. I would also like to thank my classmates in ICME for their friendship.
- ICME has been a great place to spend the past 5 years. Margot, Indira, Emily, and Antoinette have kept ICME running smoothly, for which I am very grateful.
- Microsoft Research and Technicolor for hosting me in the summers. In particular, I want to thank my mentors Ran Gilad-Bachrach, Emre Kiciman, Nadia Fawaz, and Brano Kveton for making my summers enjoyable. I would also like to thank the numerous friends I met at Microsoft Research and Technicolor.
- I would like to thank my parents for their unconditional love and support!

Contents

Abstract

Acknowledgements

1 Introduction **2**

I Selective Inference **4**

2 Selective Inference for the Lasso **5**

2.1 Introduction 5

2.1.1 The Lasso 6

2.1.2 Related Work 6

2.1.3 Outline of Paper 8

2.2 Preliminaries 9

2.3 Characterizing Selection for the Lasso 10

2.4 A Pivot for Gaussian Vectors Subject to Affine Constraints 12

2.4.1 Adaptive choice of η 15

2.5 Application to Inference for the Lasso 17

2.5.1 Confidence Intervals for the Active Variables 18

2.5.2 Testing the Lasso-Selected Model 20

2.6 Data Example 24

2.7 Minimal Post-Selection Inference 25

2.8 Extensions 29

2.8.1 Elastic net 29

2.8.2 Alternative norms as test statistics 29

2.8.3 Estimation of σ^2 . . . 30
2.8.4 Composite Null Hypotheses . . . 31
2.8.5 How long a lasso should you use? . . . 32
2.9 Conclusion . . . 34
2.10 Appendix . . . 35
2.10.1 Monotonicity of F . . . 35
2.11 Lasso Screening Property . . . 35

3 Condition-on-Selection Method **38**
3.1 Formalism . . . 38
3.2 Marginal Screening, Orthogonal Matching Pursuit, and other Variable Selection methods . . . 41
3.2.1 Marginal Screening . . . 41
3.2.2 Marginal screening + Lasso . . . 41
3.2.3 Orthogonal Matching Pursuit . . . 42
3.2.4 Nonnegative Least Squares . . . 43
3.2.5 Logistic regression with Screening . . . 44
3.3 General method for Selective inference . . . 45
3.3.1 Computational Algorithm for arbitrary selection algorithms . . . 47
3.4 Inference in the full model . . . 48
3.4.1 False Discovery Rate . . . 49
3.4.2 Intervals for coefficients in full model . . . 50
3.5 Selective Inference for the Knockoff Filter . . . 56

II Learning Mixed Graphical Models **59**

4 Learning Mixed Graphical Models **60**
4.1 Introduction . . . 60
4.2 Mixed Graphical Model . . . 61
4.2.1 Related work on mixed graphical models . . . 64
4.3 Parameter Estimation: Maximum Likelihood and Pseudolikelihood . . . 66
4.3.1 Pseudolikelihood . . . 66
4.3.2 Separate node-wise regression . . . 67

4.4 Conditional Independence and Penalty Terms 68
4.5 Calibrated regularizers 70
4.6 Model Selection Consistency 73
4.7 Optimization Algorithms 74
4.7.1 Proximal Gradient 74
4.7.2 Proximal Newton Algorithms 75
4.7.3 Path Algorithm 76
4.8 Conditional Model 77
4.9 Experimental Results 78
4.9.1 Synthetic Experiments 78
4.9.2 Survey Experiments 78
4.10 Conclusion 85
4.10.1 Proof of Convexity 87
4.10.2 Sampling From The Joint Distribution 87
4.10.3 Maximum Likelihood 88
4.10.4 Choosing the Weights 90
4.10.5 Model Selection Consistency 92

List of Tables

4.1 Frequency an edge is the first selected by the group lasso regularizer. The group lasso with equal weights is highly unbalanced, as seen in row 1. The weighing scheme with the weights from (4.5.2) is very good, and selects the edges with probability close to the ideal $\frac{1}{3}$. The approximate weighing scheme of (4.5.4) is an improvement over not calibrating; however, not as good as the weights from (4.5.2). 73

List of Figures

2.1 A picture demonstrating that the set $\{Ay \leq b\}$ can be characterized by $\{\mathcal{V}^- \leq \eta^T y \leq \mathcal{V}^+\}$. Assuming $\Sigma = I$ and $||\eta||_2 = 1$, $\mathcal{V}^-$ and $\mathcal{V}^+$ are functions of $P_{\eta^\perp} y$ only, which is independent of $\eta^T y$. 16

2.2 The density of the truncated Gaussian with distribution $F_{\mu,\sigma^2}^{[v^-,v^+]}$ depends on the width of $[v^-, v^+]$ relative to σ as well as the location of μ relative to $[v^-, v^+]$. When μ is firmly inside the interval, the distribution resembles a Gaussian. As μ varies drifts outside $[v^-, v^+]$, the density begins to converge to an exponential distribution with mean inversely proportional to the distance between μ and its projection onto $[v^-, v^+]$. 17

2.3 Histogram and empirical distribution of $F_{\eta^T\mu,\ \eta^T\Sigma\eta}^{[\mathcal{V}^-,\mathcal{V}^+]}(\eta^T y)$ obtained by sampling $y \sim N(\mu, \Sigma)$ constrained to $\{Ay \leq b\}$. The distribution is very close to $\text{Unif}(0, 1)$ as shown in Theorem 2.4.2. 18

2.4 Upper and lower bounds of 90% confidence intervals based on $[a, b] = [-3\sigma, 3\sigma]$ as a function of the observation x/σ. We see that as long as the observation x/σ is roughly 0.5σ away from either boundary, the size of the intervals is comparable to an unadjusted confidence interval. 20

2.5 90% confidence intervals for $\eta_1^T \mu$ for a small ($n = 100$, $p = 50$) and a large ($n = 100$, $p = 200$) uncorrelated Gaussian design, computed over 25 simulated data sets. The true model has five non-zero coefficients, all set to 5.0, and the noise variance is 0.25. A green bar means the confidence interval covers the true value while a red bar means otherwise. 21

2.6 P-values for $H_{0,\lambda}$ at various λ values for a small ($n = 100$, $p = 50$) and a large ($n = 100$, $p = 200$) uncorrelated Gaussian design, computed over 50 simulated data sets. The true model has three non-zero coefficients, all set to 1.0, and the noise variance is 2.0. We see the p-values are $\text{Unif}(0, 1)$ when the selected model includes the truly relevant predictors (black dots) and are stochastically smaller than $\text{Unif}(0, 1)$ when the selected model omits a relevant predictor (red dots). 24

2.7 P-values for $H_{0,\lambda}$ at various λ values for a small ($n = 100$, $p = 50$) and a large ($n = 100$, $p = 200$) *correlated* ($\rho = 0.7$) Gaussian design, computed over 50 simulated data sets. The true model has three non-zero coefficients, all set to 1.0, and the noise variance is 2.0. Since the predictors are correlated, the relevant predictors are not always selected first. However, the p-values remain uniformly distributed when $H_{0,\lambda}$ is true and stochastically smaller than $\text{Unif}(0, 1)$ otherwise. 25

2.8 Inference for the four variables selected by the lasso ($\lambda = 190$) on the diabetes data set. The point estimate and adjusted confidence intervals using the approach in Section 2.5 are shown in blue. The gray show the OLS intervals, which ignore selection. The yellow lines show the intervals produced by splitting the data into two halves, forming the interval based on only half of the data. 26

2.9 A picture demonstrating the effect of taking a union over signs. The polytope in the middle corresponds to the $(\hat{M}, \hat{z}_{\hat{M}})$ that was observed and is the same polytope as in Figure 2.1. The difference is that we now consider potential (M, z_M) in addition to the one that was observed. The polytopes for the other (M, z_M) which have the same active set $\hat{M}$ are red. The conditioning set is the union of these polytopes. We see that for y to be in this union, $\eta^T y$ must be in $\bigcup_{z_M} [\mathcal{V}^-_{z_M}, \mathcal{V}^+_{z_M}]$. The key point is that all of the $\mathcal{V}^-_{z_M}$ and $\mathcal{V}^+_{z_M}$ are still functions of only $P_{\eta^\perp} y$ and so are independent of $\eta^T y$. 27

2.10 Comparison of the minimal and simple intervals as applied to the same simulated data set for two values of λ. The simulated data featured $n = 25$, $p = 50$, and 5 true non-zero coefficients; only the first 20 coefficients are shown. (We have included variables with no intervals to emphasize that inference is only on the selected variables.) We see that the simple intervals are virtually as good as the minimal intervals most of the time; the advantage of the minimal intervals is realized when the estimate is unstable and the simple intervals are very long, as in the right plot. 28

3.1 Confidence intervals for the coefficients in a design with $n = 25$, $p = 50$, and 5 non-zero coefficients. Only the first 20 coefficients are shown. The dotted line represents the true signal, and the points represent the (biased) post-selection target. The colored bars denote the intervals. 56

4.1 *Symmetric matrix represents the parameters Θ of the model. This example has $p = 3$, $q = 2$, $L_1 = 2$ and $L_2 = 3$. The red square corresponds to the continuous graphical model coefficients B and the solid red square is the scalar β_{st}. The blue square corresponds to the coefficients ρ_{sj} and the solid blue square is a vector of parameters $\rho_{sj}(\cdot)$. The orange square corresponds to the coefficients ϕ_{rj} and the solid orange square is a matrix of parameters $\phi_{rj}(\cdot, \cdot)$. The matrix is symmetric, so each parameter block appears in two of the conditional probability regressions.* . . . 70

4.2 Row 1 shows the exact weights w_g computed via Equation (4.5.2) using Monte Carlo simulation. These are the ideal weights, but they are not available in closed-form. Row 2 shows the approximate weights computed using Equation (4.5.4). As we can see, the weights are far from uniform, and the approximate weights are close to the exact weights. 71

4.3 *Figure 4.3a shows the graph used in the synthetic experiments for $p = q = 4$; the experiment actually used p=10 and q=10. Blue nodes are continuous variables, red nodes are binary variables and the orange, green and dark blue lines represent the 3 types of edges. Figure 4.3b is a plot of the probability of correct edge recovery, meaning every true edge is selected and no non-edge is selected, at a given sample size using Maximum Likelihood and Pseudolikelihood. Results are averaged over 100 trials.* 79

4.4 Timing experiments for various instances of the graph in Figure 4.3a. The number of variables range from 20 to 2000 with $n = 500$. 79

4.5 *Model selection under different training set sizes. Circle denotes the lowest test set negative log pseudolikelihood and the number in parentheses is the number of edges in that model at the lowest test negative log pseudolikelihood. The saturated model has 55 edges.* 80

4.6 *Separate Regression vs Pseudolikelihood* $n = 100$. *y-axis is the appropriate regression loss for the response variable. For low levels of regularization and at small training sizes, the pseudolikelihood seems to overfit less; this may be due to a global regularization effect from fitting the joint distribution as opposed to separate regressions.* 82

4.7 *Separate Regression vs Pseudolikelihood* $n = 10,000$. *y-axis is the appropriate regression loss for the response variable. At large sample sizes, separate regressions and pseudolikelihood perform very similarly. This is expected since this is nearing the asymptotic regime.* 83

4.8 *Conditional Model vs Generative Model at various sample sizes. y-axis is test set performance is evaluated on negative log pseudolikelihood of the conditional model. The conditional model outperforms the full generative model at except the smallest sample size* $n = 100$. 84

4.9 *Maximum Likelihood vs Pseudolikelihood. y-axis for top row is the negative log pseudolikelihood. y-axis for bottom row is the negative log likelihood. Pseudolikelihood outperforms maximum likelihood across all the experiments.* 85

Dedicated to my parents Jerry Lee and Tien-Tien Chou.

Chapter 1

Introduction

This thesis is split into two parts: selective inference and learning mixed graphical models. The contributions are summarized below:

- Selective Inference:
  - Chapter 2: This chapter studies selective inference for the lasso-selected model. We show how to construct confidence intervals for regression coefficients corresponding to variables selected by the lasso, and how to test the significance of a lasso-selected model by conditioning on the selection event of the lasso. The results of this chapter appear in Lee et al. (2013a).
  - Chapter 3: This chapter shows how the Condition-on-Selection method developed in Chapter 2 is not specific to the lasso. In Chapter 3.1, we show that controlling the conditional type 1 error implies control of the selective type 1 error, which motivates the use of the Condition-on-Selection method to control conditional type 1 error. Chapter 3.2 studies several other variable selection methods including marginal screening, orthogonal matching pursuit, and non-negative least squares with affine selection events, so we can apply the results of Chapter 2. Motivated by more complicated selection algorithms that do not simple selection events,such as the knockoff filter, SCAD/MCP regularizers, and ℓ_1-logistic regression, we develop a general algorithm that only requires a black-box evaluation of the selection algorithm in Chapter 3.3. Finally in Chapter 3.4 we study inference for the full model regression coefficients. We show a method for FDR control, and the asymptotic coverage of selective confidence intervals

in the high-dimensional regime. The results of Chapter 3.2 appear in Lee and Taylor (2014) and the rest will appear in a future publication.

- Learning Mixed Graphical Models:
  - We propose a new pairwise Markov random field that generalizes the Gaussian graphical model to include categorical variables.
  - We design a new regularizer that promotes edge sparsity in the mixed graphical model.
  - Three methods for parameter estimation are proposed: pseudoliklihood, node-wise regression, and maximum likelihood.
  - The resulting optimization problem is solved using the proximal Newton method Lee et al. (2012).
  - We use the framework of Lee et al. (2013b) to establish edge selection consistency results for the MLE and pseudolikelihood estimation methods.
  - The results of this chapter originally appeared in Lee and Hastie (2014).

Part I

Selective Inference

Chapter 2

Selective Inference for the Lasso

2.1 Introduction

As a statistical technique, linear regression is both simple and powerful. Not only does it provide estimates of the "effect" of each variable, but it also quantifies the uncertainty in those estimates, paving the way for intervals and tests of the effect size. However, in many applications, a practitioner starts with a large pool of candidate variables, such as genes or demographic features, and does not know *a priori* which are relevant. The problem is especially acute if there are more variables than observations, when it is impossible to even fit linear regression.

A practitioner might wish to use the data to select the relevant variables and then make inference on the selected variables. As an example, one might fit a linear model, observe which coefficients are significant at level α, and report $(1-\alpha)$-confidence intervals for only the significant coefficients. However, these intervals fail to take into account the randomness in the selection procedure. In particular, the intervals do not have the stated coverage once one marginalizes over the selected model.

To see this formally, assume the usual linear model

$$y = \mu + \epsilon,\ \mu = X\beta^0,\ \epsilon \sim N(0, \sigma^2 I), \tag{2.1.1}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix and $\beta^0 \in \mathbb{R}^p$. Let $\hat{M} \subset \{1, ..., p\}$ denote a (random) set of selected variables. Suppose the goal is inference about β^0_j. Then, we do not even form intervals for β^0_j when $j \notin \hat{M}$, so the first issue is to define an interval when $j \notin \hat{M}$

in order to evaluate the coverage of this procedure. There is no obvious way to do this so that the marginal coverage is $1-\alpha$. Furthermore, as $\hat{M}$ varies, the target of the ordinary least-squares (OLS) estimator $\hat{\beta}^{OLS}_{\hat{M}}$ is not β^0, but rather

$$\beta^{\star}_{\hat{M}} := X^{+}_{\hat{M}}\mu,$$

where $X^{+}_{\hat{M}}$ denotes the Moore-Penrose pseudoinverse of $X_{\hat{M}}$. We see that $X_{\hat{M}}\beta^{\star}_{\hat{M}} = P_{\hat{M}}\mu$, the projection of μ onto the columns of $X_{\hat{M}}$, so $\beta^{\star}_{\hat{M}}$ represents the coefficients in the best linear model using only the variables in $\hat{M}$. In general, $\beta^{\star}_{\hat{M},j} \neq \beta^0_j$ unless $\hat{M}$ contains the support set of β^0, i.e., $\hat{M} \supset S := \{j : \beta^0_j \neq 0\}$. Since $\hat{\beta}^{OLS}_{\hat{M},j}$ may not be estimating β^0_j at all, there is no reason to expect a confidence interval based on it to cover β^0_j. Berk et al. (2013) provide an explicit example of the non-normality of $\hat{\beta}^{OLS}_{\hat{M},j}$ in the post-selection context. In short, inference in the linear model has traditionally been incompatible with model selection.

2.1.1 The Lasso

In this paper, we focus on a particular model selection procedure, the lasso (Tibshirani, 1996), which achieves model selection by setting coefficients to zero exactly. This is accomplished by adding an ℓ_1 penalty term to the usual least-squares objective:

$$\hat{\beta} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{2.1.2}$$

where $\lambda \geq 0$ is a penalty parameter that controls the tradeoff between fit to the data and sparsity of the coefficients. However, the distribution of the lasso estimator $\hat{\beta}$ is known only in the less interesting $n \gg p$ case (Knight and Fu, 2000), and even then, only asymptotically. Inference based on the lasso estimator is still an open question.

We apply our framework for post-selection inference about $\eta^T_{\hat{M}}\mu$ to form confidence intervals for $\beta^{\star}_{\hat{M},j}$ and to test whether the the fitted model captures all relevant signal variables.

2.1.2 Related Work

Most of the theoretical work on fitting high-dimensional linear models focuses on *consistency*. The flavor of these results is that under certain assumptions on X, the lasso fit $\hat{\beta}$

is close to the unknown β^0 (Negahban et al., 2012) and selects the correct model (Zhao and Yu, 2006; Wainwright, 2009). A comprehensive survey of the literature can be found in Bühlmann and van de Geer (2011).

There is also some recent work on obtaining confidence intervals and significance testing for penalized M-estimators such as the lasso. One class of methods uses sample splitting or subsampling to obtain confidence intervals and p-values. Recently, Meinshausen and Bühlmann (2010) proposed *stability selection* as a general technique designed to improve the performance of a variable selection algorithm. The basic idea is, instead of performing variable selection on the whole data set, to perform variable selection on random subsamples of the data of size $\binom{n}{2}$ and include the variables that are selected most often on the subsamples.

A separate line of work establishes the asymptotic normality of a corrected estimator obtained by "inverting" the KKT conditions (van de Geer et al., 2013; Zhang and Zhang, 2014; Javanmard and Montanari, 2013). The corrected estimator $\hat{b}$ usually has the form

$$\hat{b} = \hat{\beta} + \lambda \Theta \hat{z},$$

where $\hat{z}$ is a subgradient of the penalty at $\hat{\beta}$ and Θ is an approximate inverse to the Gram matrix $X^T X$. This approach is very general and easily handles M-estimators that minimize the sum of a smooth convex loss and a convex penalty. The two main drawbacks to this approach are:

1. the confidence intervals are valid only when the M-estimator is consistent
2. obtaining Θ is usually much more expensive than obtaining $\hat{\beta}$.

Most closely related to our work is the pathwise signficance testing framework laid out in Lockhart et al. (2014). They establish a test for whether a newly added coefficient is a relevant variable. This method only allows for testing at λ that are LARS knot values. This is a considerable restriction, since the lasso is often not solved with the LARS algorithm. Furthermore, the test is asymptotic, makes strong assumptions on X, and the weak convergence assumes that all relevant variables are already included in the model. They do not discuss forming confidence intervals for the selected variables. Section 2.5.2 establishes a nonasymptotic test for the same null hypothesis, while only assuming X is in general position.

In contrast, we provide a test that is exact, allows for arbitrary λ, and arbitrary design matrix X. By extension, we do not make any assumptions on n and p, and do not require the lasso to be a consistent estimator of β^0. Furthermore, the computational expense to conduct our test is negligible compared to the cost of obtaining the lasso solution.

Like all of the preceding works, our test assumes that the noise variance σ^2 is known or can be estimated. In the low-dimensional setting $p \ll n$, σ^2 can be estimated from the residual sum-of-squares of the saturated model. Strategies in high dimensions are discussed in Fan et al. (2012) and Reid et al. (2013). In Section 2.8, we also provide a strategy for estimating σ^2 based on the framework we develop.

2.1.3 Outline of Paper

We begin by defining several important quantities related to the lasso in Section 2.2; most notably, we define the selected model $\hat{M}$ in terms of the active set of the lasso solution. Section 2.3 provides an alternative characterization of the selection procedure for the lasso in terms of affine constraints on y, i.e., $Ay \leq b$. Therefore, the distribution of y conditional on the selected model is the distribution of a Gaussian vector conditional on its being in a polytope. In Section 2.4, we generalize and show that for $y \sim N(\mu, \Sigma)$, the distribution of $\eta^T y \mid Ay \leq b$ is roughly a truncated Gaussian random variable, and derive a pivot for $\eta^T \mu$. In Section 2.5, we specialize again to the lasso, deriving confidence intervals for $\beta^{\star}_{\hat{M},j}$ and hypothesis tests of the selected model as special cases of $\eta^T \mu$. Section 2.6 presents an example of these methods applied to a dataset.

In Section 2.7, we consider a refinement that produces narrower confidence intervals. Finally, Section 2.8 collects a number extensions of the framework. In particular, we demonstrate:

- modifications needed for the elastic net (Zou and Hastie, 2005).
- different norms as test statistics for the "goodness of fit" test discussed in Section 2.5.
- estimation of σ^2 based on fitting the lasso with a sufficiently small λ.
- composite null hypotheses.
- fitting the lasso for a sequence of λ values and its effect on our basic tests and intervals.

2.2 Preliminaries

Necessary and sufficient conditions for $(\hat{\beta}, \hat{z})$ to be solutions to the lasso problem (2.1.2) are the Karush-Kuhn-Tucker (KKT) conditions:

$$X^T(X\hat{\beta} - y) + \lambda\hat{z} = 0, \tag{2.2.1}$$

$$\hat{z}_i \in \begin{cases} \text{sign}(\hat{\beta}_i) & \text{if } \hat{\beta}_i \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_i = 0 \end{cases}. \tag{2.2.2}$$

where $\hat{z} := \partial||\cdot||_1(\hat{\beta})$ denotes the subgradient of the ℓ_1 norm at $\hat{\beta}$. We consider the *active set* (Tibshirani, 2013)

$$\hat{M} = \{i \in \{1, \ldots, p\} : |\hat{z}_i| = 1\}, \tag{2.2.3}$$

so-named because by examining only the rows corresponding to $\hat{M}$ in (2.2.1), we obtain the relation

$$X_{\hat{M}}^T(y - X\hat{\beta}) = -\lambda\hat{z}_{\hat{M}},$$

where $X_{\hat{M}}$ is the submatrix of X consisting of the columns in $\hat{M}$. Hence

$$|X_{\hat{M}}^T(y - X\hat{\beta})| = \lambda,$$

i.e. the variables in this set have equal (absolute) correlation with the residual $y - X\hat{\beta}$. Since $\hat{z}_i \in \{-1, 1\}$ for any $\hat{\beta}_i \neq 0$, all variables with non-zero coefficients are contained in the active set.

Recall that we are interested in inference for $\eta^T\mu$ in the model (2.1.1) for some direction $\eta = \eta_{\hat{M}} \in \mathbb{R}^n$, which is allowed to depend on the selected variables $\hat{M}$. In most applications, we will assume $\mu = X\beta^0$, although our results hold even if the linear model is not correctly specified.

A natural estimate for $\eta^T\mu$ is $\eta^T y$. As mentioned previously, we allow $\eta = \eta_{\hat{M}}$ to depend on the random selection procedure, so our goal is post-selection inference based on

$$\eta^T y \mid \{\hat{M} = M\}.$$

For reasons that will become clear, a more tractable quantity is the distribution conditional

on both the selected variables and their signs

$$\eta^T y \mid \{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}.$$

Note that confidence intervals and hypothesis tests that are valid conditional on the finer partition $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$ will also be valid for $\{\hat{M} = M\}$, by summing over the possible signs z_M:

$$\mathbb{P}(\,\cdot \,\big|\, \hat{M} = M) = \sum_{z_M} \mathbb{P}(\,\cdot \,\big|\, (\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M))\; \mathbb{P}(\hat{z}_{\hat{M}} = z_M \,\big|\, \hat{M} = M).$$

From this, it is clear that controlling $\mathbb{P}(\,\cdot \,\big|\, (\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M))$ to be, say, less than α (as in the case of hypothesis testing) will ensure $\mathbb{P}(\,\cdot \,\big|\, \hat{M} = M) \leq \alpha$.

It may not be obvious yet why we condition on $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$ instead of $\{\hat{M} = M\}$. In the next section, we show that the former can be restated in terms of affine constraints on y, i.e., $\{Ay \leq b\}$. We revisit the problem of conditioning only on $\{\hat{M} = M\}$ in Section 2.7.

2.3 Characterizing Selection for the Lasso

Recall from the previous section that our goal is inference conditional on $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$. In this section, we show that this *selection event* can be rewritten in terms of affine constraints on y, i.e.,

$$\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\} = \{A(M, z_M)y \leq b(M, z_M)\}$$

for a suitable matrix $A(M, z_M)$ and vector $b(M, z_M)$. Therefore, the conditional distribution $y \mid \{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$ is simply $y \,\big|\, \{A(M, z_M)y \leq b(M, z_M)\}$. This key theorem follows from two intermediate results.

Lemma 2.3.1. *Without loss of generality, assume the columns of X are in general position. Let $M \subset \{1, \ldots, p\}$ and $z_M \in \{-1, 1\}^{|M|}$ be a candidate set of variables and signs,*

respectively. Define

$$U = U(M, z_M) := (X_M^T X_M)^{-1}(X_M^T y - \lambda z_M) \tag{2.3.1}$$

$$W = W(M, z_M) := X_{-M}^T (X_M^T)^+ z_M + \frac{1}{\lambda} X_{-M}^T (I - P_M) y. \tag{2.3.2}$$

Then the selection procedure can be rewritten in terms of U and W as:

$$\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\} = \{\text{sign}(U(M, z_M)) = z_M, \|W(M, z_M)\|_\infty < 1\} \tag{2.3.3}$$

Proof. First, we rewrite the KKT conditions (2.2.1) and (2.2.2) by partitioning them according to the active set $\hat{M}$:

$$X_{\hat{M}}^T (X_{\hat{M}} \hat{\beta}_{\hat{M}} - y) + \lambda \hat{z}_{\hat{M}} = 0$$
$$X_{-\hat{M}}^T (X_{\hat{M}} \hat{\beta}_{\hat{M}} - y) + \lambda \hat{z}_{-\hat{M}} = 0$$
$$\text{sign}(\hat{\beta}_{\hat{M}}) = \hat{z}_{\hat{M}},\ \hat{z}_{-\hat{M}} \in (-1, 1).$$

Since the KKT conditions are necessary and sufficient for a solution, we obtain that $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$ if and only if there exist U and W satisfying:

$$X_M^T (X_M U - y) + \lambda z_M = 0 \tag{2.3.4}$$

$$X_{-M}^T (X_M U - y) + \lambda W = 0 \tag{2.3.5}$$

$$\text{sign}(U) = z_M,\ W \in (-1, 1). \tag{2.3.6}$$

Solving (2.3.4) and (2.3.5) for U and W yields the formulas (2.3.1) and (2.3.2). Finally, the requirement that U and W satisfy (2.3.6) yields (2.3.3). □

Lemma 2.3.1 is remarkable because it says that the selection event $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$ is equivalent to affine constraints on y. To see this, note that both U and W are affine functions of y, so $\{\text{sign}(U) = z_M, \|W\|_\infty < 1\}$ can be written as affine constraints $\{A(M, z_M)y \leq b(M, z_M)\}$. The following proposition provides explicit formulas for A and b.

Proposition 2.3.2. *Let U and W be defined as in (2.3.1) and (2.3.2). Then:*

$$\{\text{sign}(U) = z_M, \|W\|_\infty < 1\} = \left\{ \begin{pmatrix} A_0(M, z_M) \\ A_1(M, z_M) \end{pmatrix} y < \begin{pmatrix} b_0(M, z_M) \\ b_1(M, z_M) \end{pmatrix} \right\} \tag{2.3.7}$$

where A_0, b_0 encode the "inactive" constraints $\{\|W\|_\infty < 1\}$, and A_1, b_1 encode the "active" constraints $\{\text{sign}(U) = z_M\}$. These matrices have the explicit forms:

$$A_0(M, z_M) = \frac{1}{\lambda}\begin{pmatrix} X_{-M}^T(I - P_M) \\ -X_{-M}^T(I - P_M) \end{pmatrix} \qquad b_0(M, z_M) = \begin{pmatrix} \mathbf{1} - X_{-M}^T(X_M^T)^+ z_M \\ \mathbf{1} + X_{-M}^T(X_M^T)^+ z_M \end{pmatrix}$$
$$A_1(M, z_M) = -\operatorname{diag}(z_M)(X_M^T X_M)^{-1} X_M^T \qquad b_1(M, z_M) = -\lambda \operatorname{diag}(z_M)(X_M^T X_M)^{-1} z_M$$

Proof. First, we write

$$\{\text{sign}(U) = z_M\} = \{\operatorname{diag}(z_M)U > 0\}.$$

From here, it is straightforward to derive the above expressions from the definitions of U and W given in (2.3.1) and (2.3.2). □

Combining Lemma 2.3.1 with Proposition 2.3.2, we obtain the following.

Theorem 2.3.3. *The selection procedure can be rewritten in terms of affine constraints on y:*

$$\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\} = \{A(M, z_M)y \le b(M, z_M)\}.$$

To summarize, we have shown that in order to understand the distribution of $y \sim N(\mu, \Sigma)$ conditional on the selection procedure $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$, it suffices to study the distribution of y conditional on being in the polytope $\{Ay \le b\}$. The next section derives a pivot for $\eta^T\mu$ for such distributions, which will be useful for constructing confidence intervals and hypothesis tests in Section 2.5.

2.4 A Pivot for Gaussian Vectors Subject to Affine Constraints

The distribution of a Gaussian vector $y \sim N(\mu, \Sigma)$ conditional on affine constraints $\{Ay \le b\}$, while explicit, still involves the intractable normalizing constant $\mathbb{P}(Ay \le b)$. In this section, we derive a pivot for one dimensional projections of μ (i.e., $\eta^T\mu$), which will allow us to form tests and intervals for $\eta^T\mu$. That pivot is $F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y)$, which has a $\text{Unif}(0, 1)$ distribution conditional on $Ay \le b$.

The key to deriving this pivot is the following lemma:

Lemma 2.4.1. *The conditioning set can be rewritten in terms of $\eta^T y$ as follows:*

$$\{Ay \leq b\} = \{\mathcal{V}^-(y) \leq \eta^T y \leq \mathcal{V}^+(y), \mathcal{V}^0(y) \geq 0\}$$

where

$$a = \frac{A\Sigma\eta}{\eta^T\Sigma\eta} \tag{2.4.1}$$

$$\mathcal{V}^- = \mathcal{V}^-(y) = \max_{j:\ a_j<0} \frac{b_j - (Ay)_j + a_j\eta^T y}{a_j} \tag{2.4.2}$$

$$\mathcal{V}^+ = \mathcal{V}^+(y) = \min_{j:\ a_j>0} \frac{b_j - (Ay)_j + a_j\eta^T y}{a_j}. \tag{2.4.3}$$

$$\mathcal{V}^0 = \mathcal{V}^0(y) = \min_{j:\ a_j=0} b_j - (Ay)_j \tag{2.4.4}$$

Furthermore, $(\mathcal{V}^+, \mathcal{V}^-, \mathcal{V}^0)$ is independent of $\eta^T y$. Then, $\eta^T y$ conditioned on $Ay \leq b$ and $(\mathcal{V}^+(y), \mathcal{V}^-(y)) = (v^+, v^-)$, has a truncated normal distribution, i.e.

$$\eta^T y \,\big|\, \{\, Ay \leq b, \mathcal{V}^+(y) = v^+, \mathcal{V}^-(y) = v^-\} \sim TN(\eta^T\mu, \eta^T\Sigma\eta, v^-, v^+).$$

However, before stating the proof of this lemma, we show how it is used to obtain our main result.

Theorem 2.4.2. *Let $F^{[a,b]}_{\mu,\sigma^2}$ denote the CDF of a $N(\mu, \sigma^2)$ random variable truncated to the interval $[a, b]$, i.e.:*

$$F^{[a,b]}_{\mu,\sigma^2}(x) = \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)} \tag{2.4.5}$$

where Φ is the CDF of a $N(0, 1)$ random variable. Then $F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y)$ is a pivotal quantity, conditional on $\{Ay \leq b\}$:

$$F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y) \;\big|\; \{Ay \leq b\} \sim \text{Unif}(0, 1) \tag{2.4.6}$$

where $\mathcal{V}^-$ and $\mathcal{V}^+$ are defined in (2.4.2) and (2.4.3).

Proof. By Lemma 2.4.1, $\eta^T y \,\big|\, \{\, Ay \leq b, \mathcal{V}^+(y) = v^+, \mathcal{V}^-(y) = v^-\} \sim TN(\eta^T\mu, \eta^T\Sigma\eta, v^-, v^+)$. We apply the CDF transform to deduce

$$F^{[v^-,v^+]}_{\eta^T\mu,\eta^T\Sigma\eta}(\eta^T y) \,\big|\, \{\, Ay \leq b, \mathcal{V}^+(y) = v^+, \mathcal{V}^-(y) = v^-\}$$

is uniformly distributed. By integrating over $(\mathcal{V}^+(y) = v^+, \mathcal{V}^-(y) = v^-)$, we conclude $F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y) \mid \{Ay \le b\} \sim \text{Unif}(0,1)$. Let $G(v^+, v^-) = \mathbb{P}(\mathcal{V}^+ \le v^+, \mathcal{V}^- \le v^- \mid Ay \le b)$.

$$\begin{aligned}
&\mathbb{P}\left(F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y) \le s \;\middle|\; Ay \le b\right) \\
&= \int \mathbb{P}\left(F^{[v^-,v^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y) \le s \;\middle|\; Ay \le b, \mathcal{V}^+(y) = v^+, \mathcal{V}^-(y) = v^-\right) \\
&\qquad dG(v^+, v^-) \\
&= \int s \; dG(v^+, v^-) \\
&= s.
\end{aligned}$$

□

We now prove Lemma 2.4.1.

Proof. The linear constraints $Ay \le b$ are equivalent to

$$Ay - \mathbf{E}[Ay \mid \eta^T y] \le b - \mathbf{E}[Ay \mid \eta^T y]. \tag{2.4.7}$$

Since conditional expectation has the form

$$\mathbf{E}[Ay \mid \eta^T y] = A\mu + a(\eta^T y - \eta^T \mu),\ a = \frac{A\Sigma\eta}{\eta^T\Sigma\eta},$$

(2.4.7) simplifies to $Ay - b - a\eta^T y \le -a\eta^T y$. Rearranging, we obtain

$$\begin{aligned}
\eta^T y &\ge \frac{1}{a_j}(b_j - (Ay)_j + a_j\eta^T y) && a_j < 0 \\
\eta^T y &\le \frac{1}{a_j}(b_j - (Ay)_j + a_j\eta^T y) && a_j > 0 \\
0 &\le b_j - (Ay)_j && a_j = 0.
\end{aligned}$$

We take the max of the lower bounds and min of the upper bounds to deduce

$$\underbrace{\max_{j:a_j<0} \frac{1}{a_j}(b_j - (Ay)_j + a_j\eta^T y)}_{\mathcal{V}^-(y)} \le \eta^T y \le \underbrace{\min_{j:a_j>0} \frac{1}{a_j}(b_j - (Ay)_j + a_j\eta^T y)}_{\mathcal{V}^+(y)}.$$

Since y is normal, $b_j - (Ay)_j + a_j\eta^T y,\ j = 1, \ldots, m$ are independent of $\eta^T y$. Hence

$(\mathcal{V}^+(y), \mathcal{V}^-(y), \mathcal{V}^0(y))$ are also independent of $\eta^T y$.

To complete the proof, we must show $\eta^T y$ given $Ay \leq b$, $(\mathcal{V}^+(y), \mathcal{V}^-(y)) = (v^+, v^-)$ is truncated normal.

$$\begin{aligned}
&\mathbb{P}\left(\eta^T y \leq s \,\middle|\, Ay \leq b, \mathcal{V}^+(y) = v^+, \mathcal{V}^-(y) = v^-,\right) \\
&= \mathbb{P}\left(\eta^T y \leq s \,\middle|\, v^- \leq \eta^T y \leq v^+, \mathcal{V}^+(y) = v^+, \mathcal{V}^-(y) = v^-, \mathcal{V}^0(y) \geq 0\right) \\
&= \frac{\mathbb{P}\left(\eta^T y \leq s, v^- \leq \eta^T y \leq v^+ \middle| \mathcal{V}^+(y) = v^+, \mathcal{V}^-(y) = v^-, \mathcal{V}^0(y) \geq 0\right)}{\mathbb{P}\left(v^- \leq \eta^T y \leq v^+ \middle| \mathcal{V}^+(y) = v^+, \mathcal{V}^-(y) = v^-, \mathcal{V}^0(y) \geq 0\right)} \\
&= \frac{\mathbb{P}\left(\eta^T y \leq s, v^- \leq \eta^T y \leq v^+\right)}{\mathbb{P}\left(v^- \leq \eta^T y \leq v^+\right)} = \mathbb{P}\left(\eta^T y \leq s \mid v^- \leq \eta^T y \leq v^+\right)
\end{aligned}$$

where the second to last equality follows from the independence of $(\mathcal{V}^+, \mathcal{V}^-, \mathcal{V}^0)$ and $\eta^T y$. This is the CDF of a truncated normal. □

Although the proof of Lemma 2.4.1 is elementary, the geometric picture gives more intuition as to why $\mathcal{V}^+$ and $\mathcal{V}^-$ are independent of $\eta^T y$. Without loss of generality, we assume $||\eta||_2 = 1$ and $y \sim N(\mu, I)$ (since otherwise we could replace y by $\Sigma^{-\frac{1}{2}} y$). Now we can decompose y into two independent components, a 1-dimensional component $\eta^T y$ and an $(n-1)$-dimensional component orthogonal to η:

$$y = \eta^T y + P_{\eta^\perp} y.$$

The case of $n = 2$ is illustrated in Figure 2.1. $\mathcal{V}^-$ and $\mathcal{V}^+$ are independent of $\eta^T y$, since they are functions of $P_{\eta^\perp}$ only, which is independent of $\eta^T y$.

In Figure 2.2, we plot the density of the truncated Gaussian, noting that its shape depends on the location of μ relative to $[a, b]$ as well as the width relative to σ.

2.4.1 Adaptive choice of η

For the applications to forming confidence intervals and significance testing, we will need choices of η that are adaptive, or dependent on y. We will restrict ourselves to functions η that are functions of the partition, $\eta(y) = f(\hat{M}(y))$. This choice of functions includes $\eta(y) = X^{T+}_{\hat{M}(y)} e_j$ which is used for forming confidence intervals in Section 2.5.

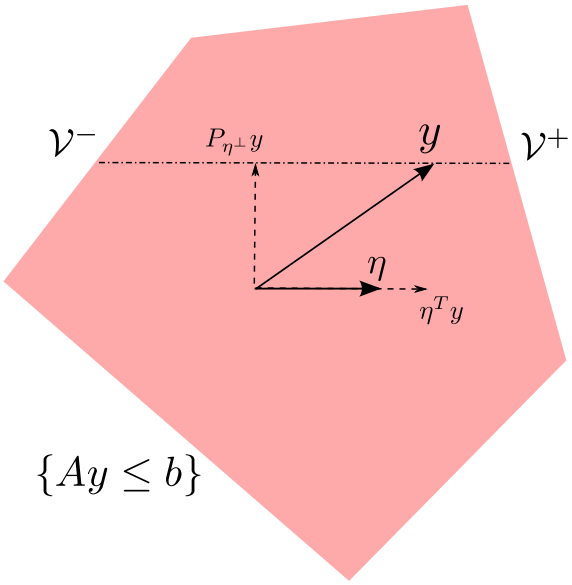


Figure 2.1: A picture demonstrating that the set $\{Ay \leq b\}$ can be characterized by $\{\mathcal{V}^- \leq \eta^T y \leq \mathcal{V}^+\}$. Assuming $\Sigma = I$ and $||\eta||_2 = 1$, $\mathcal{V}^-$ and $\mathcal{V}^+$ are functions of $P_{\eta^\perp} y$ only, which is independent of $\eta^T y$.

Theorem 2.4.3. *Let $\eta : \mathbf{R}^n \to \mathbf{R}^n$ be a function of the form $\eta(y) = f(\hat{M}(y))$, then*

$$F^{[\mathcal{V}^-(y),\mathcal{V}^+(y)]}_{\eta(y)^T\mu,\ \eta(y)^T\Sigma\eta(y)} \left(\eta(y)^T y\right) \sim \text{Unif}(0,1).$$

Proof. We can expand F with respect to the partition,

$$\begin{aligned}
&\mathbb{P}\left(F^{[\mathcal{V}^-(y),\mathcal{V}^+(y)]}_{\eta(y)^T\mu,\ \eta(y)^T\Sigma\eta(y)} \left(\eta(y)^T y\right) \leq t\right) = \sum_{(M,s)} \mathbb{P}\left(F^{[\mathcal{V}^-(y),\mathcal{V}^+(y)]}_{\eta(y)^T\mu,\ \eta(y)^T\Sigma\eta(y)} \left(\eta(y)^T y\right) \leq t, \hat{M}(y) = M\right) \\
&= \sum_{(M,s)} \mathbb{P}\left(F^{[\mathcal{V}^-(y),\mathcal{V}^+(y)]}_{\eta(y)^T\mu,\ \eta(y)^T\Sigma\eta(y)} \left(\eta(y)^T y\right) \leq t \middle| \hat{M}(y) = M\right) \mathbb{P}\left(\hat{M}(y) = M\right) \\
&= \sum_{(M,s)} \mathbb{P}\left(F^{[\mathcal{V}^-(y),\mathcal{V}^+(y)]}_{f(M)^T\mu,\ f(M)^T\Sigma f(M)} \left(f(M)^T y\right) \leq t \middle| \hat{M}(y) = M\right) \mathbb{P}\left(\hat{M}(y) = M\right)
\end{aligned}$$

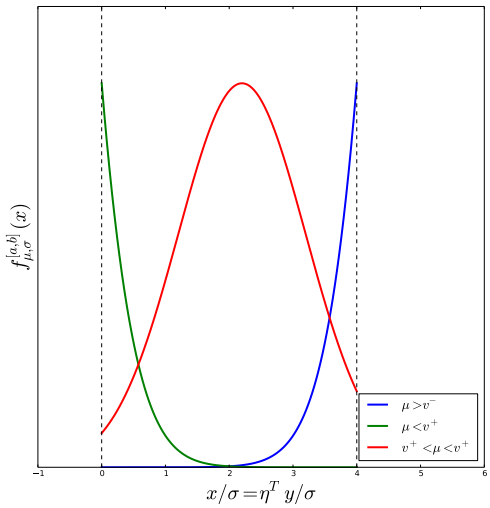


Figure 2.2: The density of the truncated Gaussian with distribution $F^{[v^-,v^+]}_{\mu,\sigma^2}$ depends on the width of $[v^-, v^+]$ relative to σ as well as the location of μ relative to $[v^-, v^+]$. When μ is firmly inside the interval, the distribution resembles a Gaussian. As μ varies drifts outside $[v^-, v^+]$, the density begins to converge to an exponential distribution with mean inversely proportional to the distance between μ and its projection onto $[v^-, v^+]$.

Using Theorem 2.4.2, $\mathbb{P}\left(F^{[\mathcal{V}^-(y),\mathcal{V}^+(y)]}_{f(M)^T\mu,\ f(M)^T\Sigma f(M)}\left(f(M)^T y\right) \leq t \middle| \hat{M}(y) = M\right) = t$. Thus

$$\begin{aligned}
\mathbb{P}\left(F^{[\mathcal{V}^-(y),\mathcal{V}^+(y)]}_{\eta(y)^T\mu,\ \eta(y)^T\Sigma\eta(y)}\left(\eta(y)^T y\right)\right) \leq t\Big) &= \sum_{(M,s)} t\mathbb{P}\left(\hat{M}(y) = M\right) \\
&= t\sum_{(M,s)} \mathbb{P}\left(\hat{M}(y) = M\right) \\
&= t.
\end{aligned}$$

This shows that $F^{[\mathcal{V}^-(y),\mathcal{V}^+(y)]}_{\eta(y)^T\mu,\ \eta(y)^T\Sigma\eta(y)}\left(\eta(y)^T y\right) \sim \text{Unif}(0,1)$. □

2.5 Application to Inference for the Lasso

In this section, we apply the theory developed in in Sections 2.3 and 2.4 to the lasso. In particular, we will construct confidence intervals for the active variables and test the chosen model based on the pivot developed in Section 2.4.

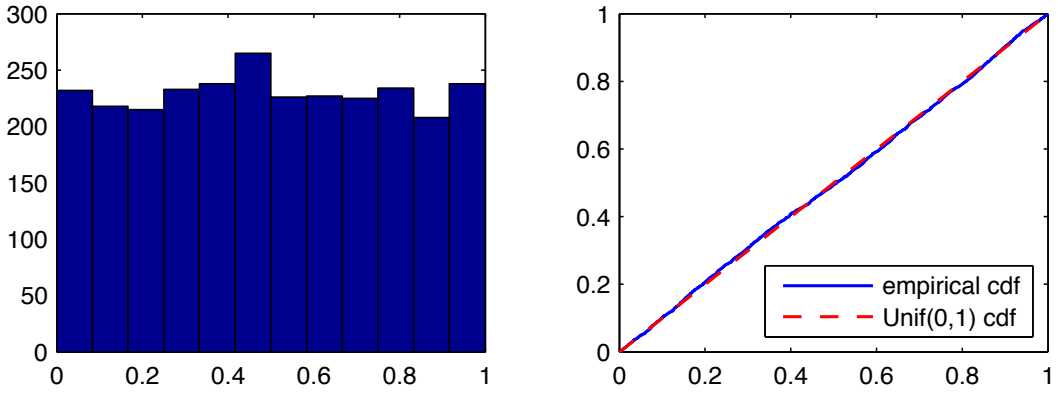


Figure 2.3: Histogram and empirical distribution of $F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y)$ obtained by sampling $y \sim N(\mu, \Sigma)$ constrained to $\{Ay \leq b\}$. The distribution is very close to $\text{Unif}(0,1)$ as shown in Theorem 2.4.2.

To summarize the developments so far, recall that our model says that $y \sim N(\mu, \sigma^2 I)$. The distribution of interest is $y \mid \{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$. By Theorem 2.3.1, this is equivalent to $y \mid \{A(M, z_M)y \leq b(M, z_M)\}$ defined in Proposition 2.3.2. Now we can apply Theorem 2.4.2 to obtain the (conditional) pivot

$$F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^T\mu,\ \sigma^2||\eta||_2^2}(\eta^T y) \ \Big|\ \{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\} \sim \text{Unif}(0,1) \tag{2.5.1}$$

for any η, where $\mathcal{V}^-$ and $\mathcal{V}^+$ are defined in (2.4.2) and (2.4.3). Note that $A(M, z_M)$ and $b(M, z_M)$ appear in this pivot through $\mathcal{V}^-$ and $\mathcal{V}^+$. This pivot will play a central role in all of the applications that follow.

2.5.1 Confidence Intervals for the Active Variables

In this section, we describe how to form confidence intervals for the components of $\beta^\star_{\hat{M}} = X^+_{\hat{M}}\mu$. If we choose

$$\eta_j = (X^T_{\hat{M}})^+ e_j, \tag{2.5.2}$$

then $\eta_j^T\mu = \beta^\star_{\hat{M},j}$, so the above framework provides a method for inference about the j^{th} variable in the model $\hat{M}$. Note that this reduces to inference about the true β^0_j if $\hat{M} \supset S := \{j : \beta^0_j \neq 0\}$, as discussed in Section 2.1. The conditions under which this holds is well known in the literature, cf. Bühlmann and van de Geer (2011).

By applying Theorem 2.4.2, we obtain the following (conditional) pivot for $\beta^\star_{\hat{M},j}$:

$$F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\beta^\star_{\hat{M},j},\ \sigma^2||\eta_j||^2}(\eta_j^T y) \mid \{(\hat{M},\hat{z}_{\hat{M}}) = (M, z_M)\} \sim \text{Unif}(0,1).$$

Note that j and η_j are both random—but only through $\hat{M}$, a quantity which is fixed after conditioning—so Theorem 2.4.2 holds even for this "random" choice of η. The obvious way to obtain an interval is to "invert" the pivot. In other words, since

$$\mathbb{P}\left(\frac{\alpha}{2} \le F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\beta^\star_{\hat{M},j},\ \sigma^2||\eta_j||^2}(\eta_j^T y) \le 1-\frac{\alpha}{2} \mid \{(\hat{M},\hat{z}_{\hat{M}}) = (M, z_M)\}\right) = \alpha,$$

one can define a $(1-\alpha)$ (conditional) confidence interval for $\beta^\star_{\hat{M},j}$ as

$$\left\{\beta^\star_{\hat{M},j} : \frac{\alpha}{2} \le F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\beta^\star_{\hat{M},j},\ \sigma^2||\eta_j||^2}(\eta_j^T y) \le 1-\frac{\alpha}{2}\right\}.$$

In fact, F is monotone decreasing in $\beta^\star_{\hat{M},j}$, so to find its endpoints, one need only solve for the root of a smooth one-dimensional function. The monotonicity is a consequence of the fact that the truncated Gaussian distribution is a natural exponential family and hence has monotone likelihood ratio in μ. The details can be found in Appendix 2.10.1.

We now formalize the above observations in the following result, an immediate consequence of Theorem 2.4.2.

Corollary 2.5.1. *Let η_j be defined as in (2.5.2), and let $L^j_\alpha = L^j_\alpha(\eta_j, \hat{M}, \hat{z}_{\hat{M}})$ and $U^j_\alpha = U^j_\alpha(\eta_j, \hat{M}, \hat{z}_{\hat{M}})$ be the (unique) values satisfying*

$$F^{[\mathcal{V}^-,\mathcal{V}^+]}_{L^j_\alpha,\ \sigma^2||\eta_j||^2}(\eta_j^T y) = 1-\alpha \qquad\qquad F^{[\mathcal{V}^-,\mathcal{V}^+]}_{U^j_\alpha,\ \sigma^2||\eta_j||^2}(\eta_j^T y) = \alpha$$

Then $[L^j_\alpha, U^j_\alpha]$ is a $(1-\alpha)$ confidence interval for $\eta_j^T\mu$, conditional on $(\hat{M},\hat{z}_{\hat{M}})$:

$$\mathbb{P}\left(\beta^\star_{\hat{M},j} \in [L^j_\alpha, U^j_\alpha] \mid \{(\hat{M},\hat{z}_{\hat{M}}) = (M, z_M)\}\right) = 1-\alpha. \tag{2.5.3}$$

The above discussion has focused on constructing intervals for a single j. If we repeat the procedure for each $j \in \hat{M}$, our intervals in fact control the false coverage rate (FCR) of Benjamini and Yekutieli (2005).

Corollary 2.5.2. *For each $j \in \hat{M}$,*

$$\mathbb{P}\left(\beta^{\star}_{\hat{M},j} \in [L^j_\alpha, U^j_\alpha]\right) = 1 - \alpha. \tag{2.5.4}$$

Furthermore, the FCR of the intervals $\left\{[L^j_\alpha, U^j_\alpha]\right\}_{j\in\hat{M}}$ is α.

If $\eta^T y$ are not near the boundaries $[\mathcal{V}^-, \mathcal{V}^+]$, then the intervals will be relatively short. This is shown in Figure 2.4. Figure 2.5 shows two simulations that demonstrate our intervals cover at the nominal rate. We leave an exhaustive study of such intervals for the lasso to future work, noting that the truncation framework described can be used to form intervals with exact coverage properties.

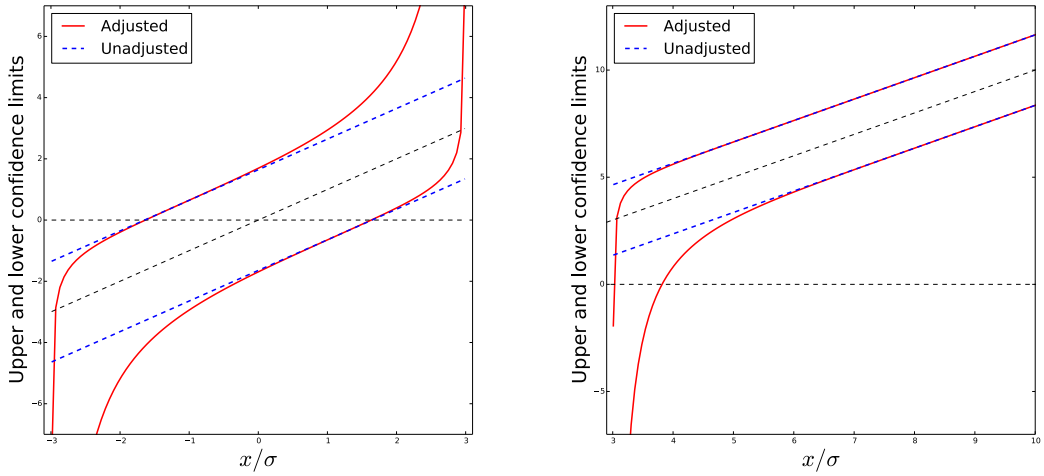


Figure 2.4: Upper and lower bounds of 90% confidence intervals based on $[a, b] = [-3\sigma, 3\sigma]$ as a function of the observation x/σ. We see that as long as the observation x/σ is roughly 0.5σ away from either boundary, the size of the intervals is comparable to an unadjusted confidence interval.

2.5.2 Testing the Lasso-Selected Model

Having observed that the lasso selected the variables $\hat{M}$, another relevant question is whether it has captured all of the signal in the model, i.e.,

$$H_0 : \beta^0_{-\hat{M}} = \mathbf{0}. \tag{2.5.5}$$

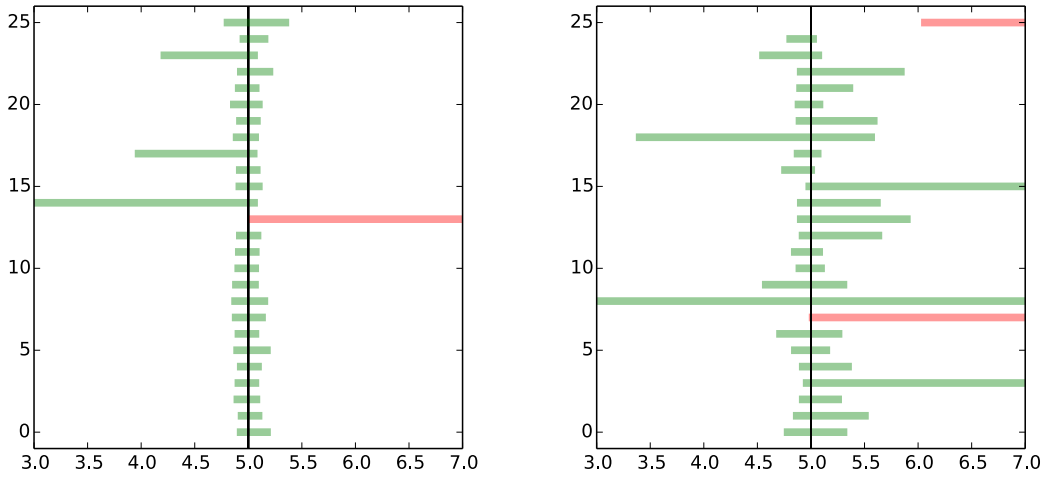


Figure 2.5: 90% confidence intervals for $\eta_1^T \mu$ for a small ($n = 100$, $p = 50$) and a large ($n = 100$, $p = 200$) uncorrelated Gaussian design, computed over 25 simulated data sets. The true model has five non-zero coefficients, all set to 5.0, and the noise variance is 0.25. A green bar means the confidence interval covers the true value while a red bar means otherwise.

We consider a slightly more general question, which does not assume the correctness of the linear model $\mu = X\beta^0$ and also takes into account whether the non-selected variables can improve the fit:

$$H_0 : X_{-\hat{M}}^T (I - P_{\hat{M}})\mu = \mathbf{0}. \tag{2.5.6}$$

This quantity is the partial correlation of the non-selected variables with μ, adjusting for the variables in $\hat{M}$. This is more general because if we assume $\mu = X\beta^0$ for some β^0 and X is full rank, then rejecting (2.5.6) implies that there exists $i \in \text{supp}(\beta^0)$ not in $\hat{M}$, so we would also reject (2.5.5).

The natural approach is to compare the observed partial correlations $X_{-M}^T (I - P_M)y$ to $\mathbf{0}$. However, the framework of Section 2.4 only allows tests of μ in a single direction η. To make use of that framework, we can choose η such that it selects the maximum magnitude of $X_{-M}^T (I - P_M)y$. In particular, this direction provides the most evidence against the null hypothesis of zero partial correlation, so if the null hypothesis cannot be rejected in this direction, it would not be rejected in any direction.

Letting $j^\star := \text{argmax}_j \, |e_j^T X_{-M}^T (I - P_M) y|$ and $s_j := \text{sign}(e_j^T X_{-M}^T (I - P_M) y)$, we set

$$\eta_{j^\star} = s_{j^\star} (I - P_M) X_{-M} e_{j^\star}, \tag{2.5.7}$$

and test $H_0 : \eta_{j^\star}^T \mu = 0$. However, the results in Section 2.4 cannot be directly applied to this setting because $j^\star$ and $s_{j^\star}$ are random variables that are not measurable with respect to $(\hat{M}, \hat{z}_{\hat{M}})$.

To resolve this issue, we propose a test conditional not only on $(\hat{M}, \hat{z}_{\hat{M}})$, but also on the index and sign of the maximizer:

$$\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M), \, (j^\star, s_{j^\star}) = (j, s)\}. \tag{2.5.8}$$

A test that is level α conditional on (2.5.8) for all (M, z_M) and (j, s) is also level α conditional on $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$.

In order to use the results of Section 2.4, we must show that (2.5.8) can be written in the form $A(M, z_M, j, s) y \leq b(M, z_M, j, s)$. This is indeed possible, and the following proposition provides an explicit construction.

Proposition 2.5.3. *Let A_0, b_0, A_1, b_1 be defined as in Proposition 2.3.2. Then:*

$$\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M), \, (j^\star, s_{j^\star}) = (j, s)\} = \left\{ \begin{pmatrix} A_0(M, z_M) \\ A_1(M, z_M) \\ A_2(M, j, s) \end{pmatrix} y < \begin{pmatrix} b_0(M, z_M) \\ b_1(M, z_M) \\ \mathbf{0} \end{pmatrix} \right\}$$

where $A_2(M, j, s)$ is defined as

$$A_2(M, j, s) = -s \begin{pmatrix} D_j(M) \\ S_j(M) \end{pmatrix} X_{-M}^T (I - P_M)$$

and D_j and S_j are $(|M| - 1) \times |M|$ operators that compute the difference and sum, respectively, of the j^{th} element with the other elements, e.g.,

$$D_1 = \begin{pmatrix} 1 & -1 & & & \\ 1 & & -1 & & \\ & & & \ddots & \\ 1 & & & & -1 \end{pmatrix} \qquad S_1 = \begin{pmatrix} 1 & 1 & & & \\ 1 & & 1 & & \\ & & & \ddots & \\ 1 & & & & 1 \end{pmatrix}.$$

Proof. The constraints $\{A_0 y < b_0\}$ and $\{A_1 y < b_1\}$ come from Proposition (2.3.2) and encode the constraints $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$. We show that the last two sets of constraints encode $\{(j^\star, s_{j^\star}) = (j, s)\}$.

Let $r := X_{-M}^T(I - P_M)y$ denote the vector of partial correlations. If $s = +1$, then $|r_j| > |r_i|$ for all $i \neq j$ if and only if $r_j - r_i > 0$ and $r_j + r_i > 0$ for all $i \neq j$. We can write this as $D_j r > 0$ and $S_j r > 0$. If $s = -1$, then the signs are flipped: $D_j r < 0$ and $S_j r < 0$. This establishes

$$\{(j^\star, s_{j^\star}) = (j, s)\} = \left\{ -s \begin{pmatrix} D_j \\ S_j \end{pmatrix} r < \mathbf{0} \right\} = \{A_2 y < \mathbf{0}\}.$$

□

Because of Proposition 2.5.3, we can now obtain the following result as a simple consequence of Theorem 2.4.2, which says that $F^{[\mathcal{V}^-, \mathcal{V}^+]}_{0, \sigma^2||\eta_{j^\star}||^2}(\eta_{j^\star}^T y) \sim \text{Unif}(0, 1)$, conditional on the set (2.5.8) and H_0. We reject when $F^{[\mathcal{V}^-, \mathcal{V}^+]}_{0, \sigma^2||\eta_j^*||^2}(\eta_{j^\star}^T y)$ is large because $F^{[\mathcal{V}^-, \mathcal{V}^+]}_{0,\ \sigma^2||\eta_j^*||^2}(\cdot)$ is monotone increasing in the argument and $\eta_{j^*}^T \mu$ is likely to be positive under the alternative.

Corollary 2.5.4. *Let H_0 and $\eta_{j^\star}$ be defined as in (2.5.7). Then, the test which rejects when*

$$\left\{ F^{[\mathcal{V}^-, \mathcal{V}^+]}_{0,\ \sigma^2||\eta_j^*||^2}(\eta_{j^\star}^T y) > 1 - \alpha \right\}$$

is level α, conditional on $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M), (j^\star, s_{j^\star}) = (j, s)\}$. That is,

$$\mathbb{P}\left(F^{[\mathcal{V}^-, \mathcal{V}^+]}_{0,\ \sigma^2||\eta_{j^\star}||^2}(\eta_{j^\star}^T y) > 1 - \alpha \;\middle|\; \{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M),\ (j^\star, s_{j^\star}) = (j, s)\} \cap H_0 \right) = \alpha.$$

In particular, since this holds for every (M, z_M, j, s), this test also controls Type I error conditional only on $(\hat{M}, \hat{z}_{\hat{M}})$, and unconditionally:

$$\mathbb{P}\left(F^{[\mathcal{V}^-, \mathcal{V}^+]}_{0,\ \sigma^2||\eta_{j^\star}||^2}(\eta_{j^\star}^T y) > 1 - \alpha \;\middle|\; \{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\} \cap H_0 \right) = \alpha$$
$$\mathbb{P}\left(F^{[\mathcal{V}^-, \mathcal{V}^+]}_{0,\ \sigma^2||\eta_{j^\star}||^2}(\eta_{j^\star}^T y) > 1 - \alpha \;\middle|\; H_0 \right) = \alpha.$$

Figures 2.6 and 2.7 show the results of four simulation studies that demonstrate that the p-values are uniformly distributed when $H_{0,\lambda}$ is true and stochastically smaller than $\text{Unif}(0, 1)$ when it is false.

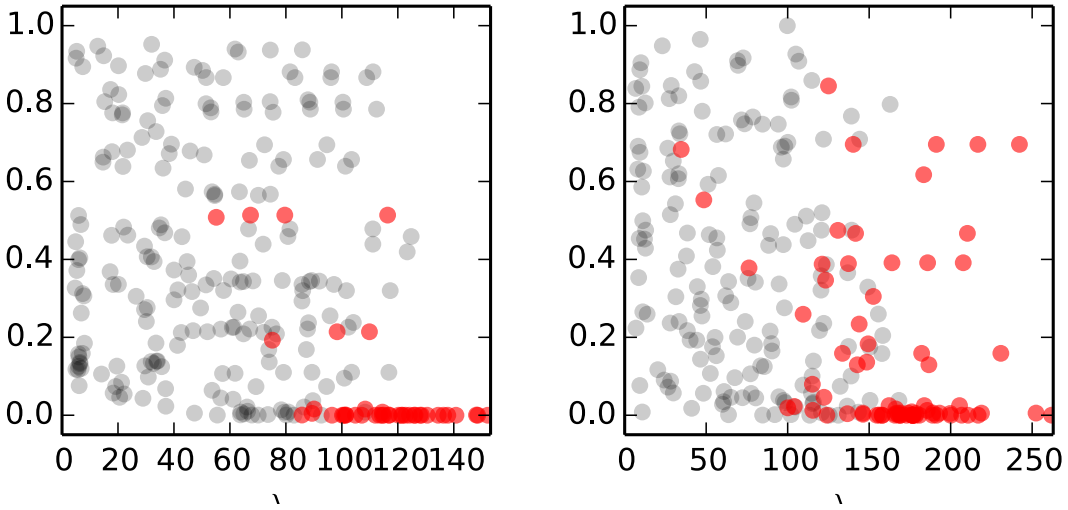


Figure 2.6: P-values for $H_{0,\lambda}$ at various λ values for a small ($n = 100$, $p = 50$) and a large ($n = 100$, $p = 200$) uncorrelated Gaussian design, computed over 50 simulated data sets. The true model has three non-zero coefficients, all set to 1.0, and the noise variance is 2.0. We see the p-values are $\text{Unif}(0, 1)$ when the selected model includes the truly relevant predictors (black dots) and are stochastically smaller than $\text{Unif}(0, 1)$ when the selected model omits a relevant predictor (red dots).

2.6 Data Example

We illustrate the application of inference for the lasso to the diabetes data set from Efron et al. (2004). First, all variables were standardized. Then, we chose λ according to the strategy in Negahban et al. (2012), $\lambda = 2\,\mathbf{E}(\|X^T\epsilon\|_\infty)$, using an estimate of σ from the full model, resulting in $\lambda \approx 190$. The lasso selected four variables: `BMI`, `BP`, `S3`, and `S5`.

The intervals are shown in Figure 2.8, alongside the unadjusted confidence intervals produced by fitting OLS to the four selected variables, ignoring the selection. The latter is not a valid confidence interval conditional on the model. Also depicted are the confidence intervals obtained by *data splitting*; that is, if one splits the n observations into two halves, then uses one half for model selection and the other for inference. This is a competitor method that also produces valid confidence intervals conditional on the model. In this case, data splitting selected the same four variables, and the confidence intervals were formed based on OLS on the half of the data set not used for model selection.

We can make two main observations from Figure 2.8.

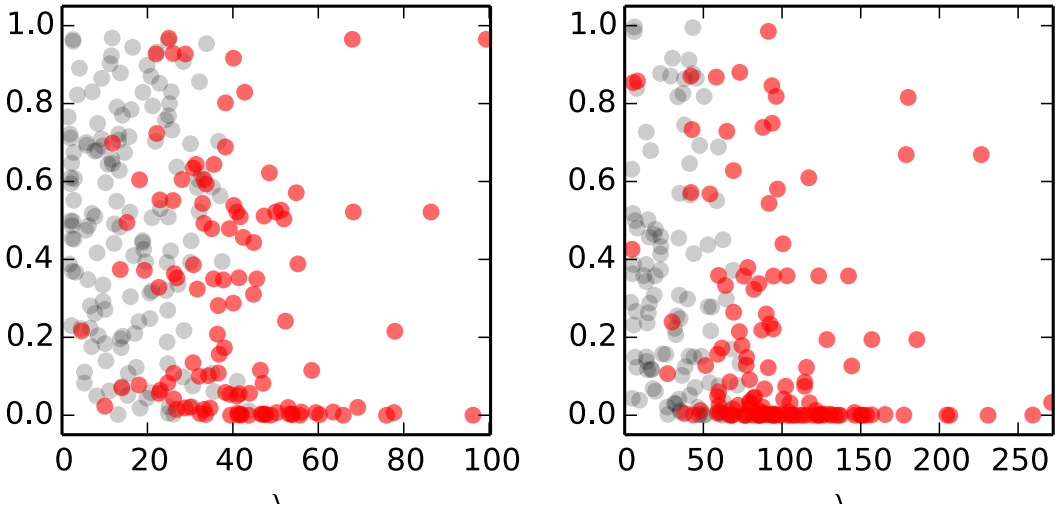


Figure 2.7: P-values for $H_{0,\lambda}$ at various λ values for a small ($n = 100$, $p = 50$) and a large ($n = 100$, $p = 200$) *correlated* ($\rho = 0.7$) Gaussian design, computed over 50 simulated data sets. The true model has three non-zero coefficients, all set to 1.0, and the noise variance is 2.0. Since the predictors are correlated, the relevant predictors are not always selected first. However, the p-values remain uniformly distributed when $H_{0,\lambda}$ is true and stochastically smaller than $\text{Unif}(0, 1)$ otherwise.

1. The adjusted intervals provided by our method essentially reproduces the OLS intervals for the strong effects, whereas data splitting results in a loss of power by roughly a factor of $\sqrt{2}$ (since only $n/2$ observations are used in the inference).

2. One variable, `S3`, which would have been deemed significant using the OLS intervals, is no longer significant after adjustment. This demonstrates that taking model selection into account can have substantive impacts on the conclusions that are made.

2.7 Minimal Post-Selection Inference

We have described how to perform post-selection inference for the lasso conditional on both the active set and signs $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$. However, recall from Section 2.1 that the goal was inference conditional solely on the model, i.e., $\{\hat{M} = M\}$. In this section, we extend our framework to this setting, which we call minimal post-selection inference because we condition on the minimal set necessary for the random η to be measurable. This results in

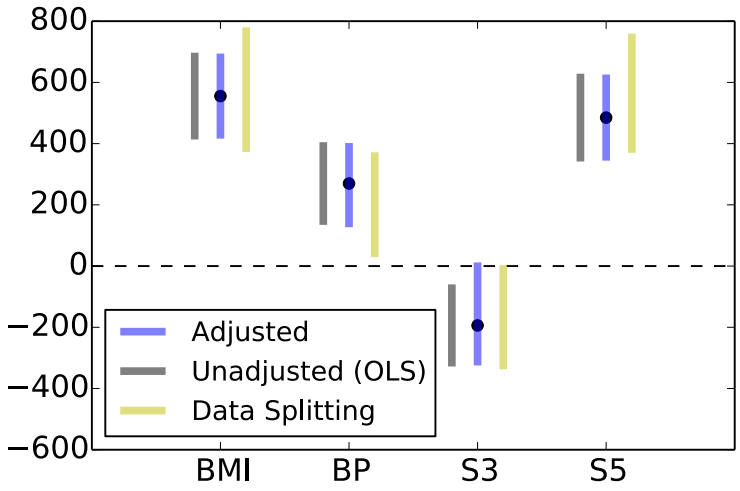


Figure 2.8: Inference for the four variables selected by the lasso ($\lambda = 190$) on the diabetes data set. The point estimate and adjusted confidence intervals using the approach in Section 2.5 are shown in blue. The gray show the OLS intervals, which ignore selection. The yellow lines show the intervals produced by splitting the data into two halves, forming the interval based on only half of the data.

more precise confidence intervals at the expense of greater computational cost.

To this end, we note that $\{\hat{M} = M\}$ is simply

$$\bigcup_{z_M \in \{-1,1\}^{|E|}} \{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\},$$

where the union is taken over all choices of signs. Therefore, the distribution of y conditioned on only the active set $\{\hat{M} = M\}$ is a Gaussian vector constrained to a union of polytopes

$$y \mid \bigcup_{z_M \in \{-1,1\}^{|E|}} \{A(M, z_M)y \leq b(M, z_M)\},$$

where $A(M, z_M)$ and $b(M, z_M)$ are given by (2.3.2).

To obtain inference about $\eta^T \mu$, we follow the arguments in Section 2.4 to obtain that this conditional distribution is equivalent to

$$\eta^T y \mid \bigcup_{z_M \in \{-1,1\}^{|E|}} \{\mathcal{V}^-_{z_M}(y) \leq \eta^T y \leq \mathcal{V}^+_{z_M}(y), \mathcal{V}^0_{z_M}(y) \geq 0\}, \tag{2.7.1}$$

where $\mathcal{V}^-_{z_M}$, $\mathcal{V}^+_{z_M}$, $\mathcal{V}^0_{z_M}$ are defined according to (2.4.2), (2.4.3), (2.11.4) with $A = A(M, z_M)$ and $b = b(M, z_M)$. Moreover, all of these quantities are still independent of $\eta^T y$, so instead of having a Gaussian truncated to a single interval $[\mathcal{V}^-, \mathcal{V}^+]$ as in Section 2.4, we now have a Gaussian truncated to the union of intervals $\bigcup_{z_M}[\mathcal{V}^-_{z_M}, \mathcal{V}^+_{z_M}]$. The geometric intuition is illustrated in Figure 2.9.

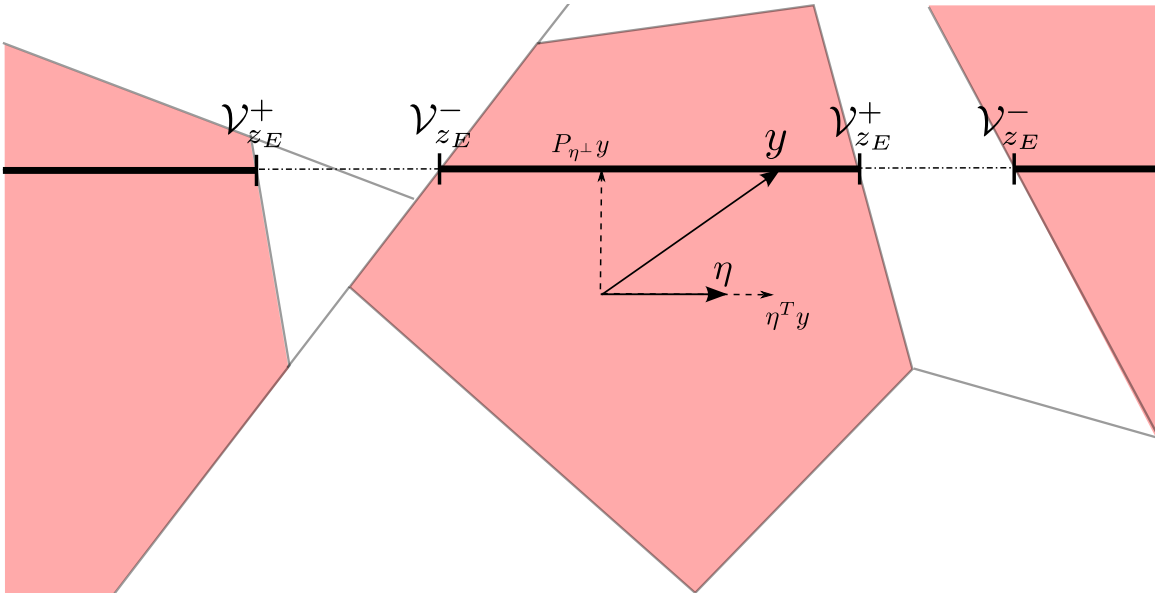


Figure 2.9: A picture demonstrating the effect of taking a union over signs. The polytope in the middle corresponds to the $(\hat{M}, \hat{z}_{\hat{M}})$ that was observed and is the same polytope as in Figure 2.1. The difference is that we now consider potential (M, z_M) in addition to the one that was observed. The polytopes for the other (M, z_M) which have the same active set $\hat{M}$ are red. The conditioning set is the union of these polytopes. We see that for y to be in this union, $\eta^T y$ must be in $\bigcup_{z_M}[\mathcal{V}^-_{z_M}, \mathcal{V}^+_{z_M}]$. The key point is that all of the $\mathcal{V}^-_{z_M}$ and $\mathcal{V}^+_{z_M}$ are still functions of only $P_{\eta^\perp} y$ and so are independent of $\eta^T y$.

Finally, the probability integral transform once again yields a pivot:

$$F^{\bigcup_{z_E}[\mathcal{V}^-_{z_E}(y), \mathcal{V}^+_{z_E}(y)]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y) \ \Big|\ \{\hat{M} = M\} \sim \text{Unif}(0, 1).$$

It is now more useful to think of the notation of F as indicating the truncation set $C \subset \mathbb{R}$:

$$F^C_{\mu,\sigma^2}(x) := \frac{\Phi((-\infty, x] \cap C)}{\Phi(C)}, \tag{2.7.2}$$

where Φ is the law of a $N(0, 1)$ random variable. We summarize these results in the following

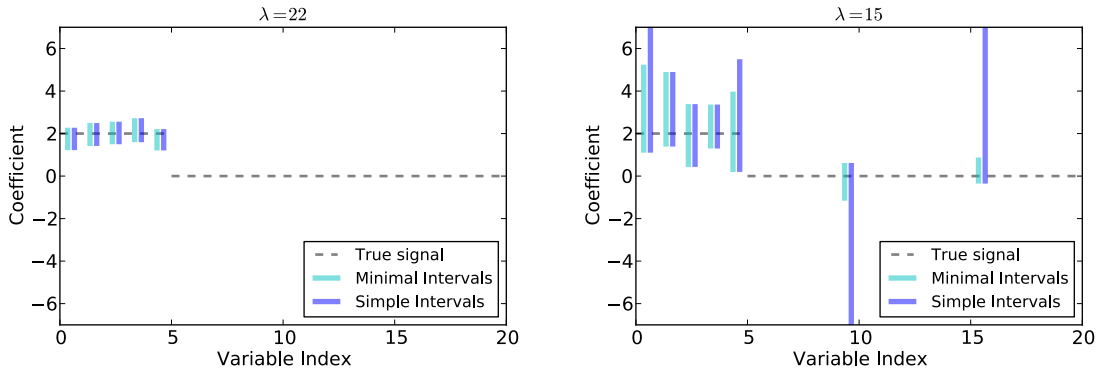


Figure 2.10: Comparison of the minimal and simple intervals as applied to the same simulated data set for two values of λ. The simulated data featured $n = 25$, $p = 50$, and 5 true non-zero coefficients; only the first 20 coefficients are shown. (We have included variables with no intervals to emphasize that inference is only on the selected variables.) We see that the simple intervals are virtually as good as the minimal intervals most of the time; the advantage of the minimal intervals is realized when the estimate is unstable and the simple intervals are very long, as in the right plot.

theorem.

Theorem 2.7.1. *Let* $F^{\bigcup_i [a_i, b_i]}_{\mu,\sigma^2}$ *be the CDF of a normal truncated to the union of intervals* $\bigcup_i [a_i, b_i]$*, i.e., given by* (2.7.2)*. Then:*

$$F^{\bigcup_{z_M} [\mathcal{V}^-_{z_M}(y), \mathcal{V}^+_{z_M}(y)]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y) \mid \{\hat{M} = M\} \sim \mathrm{Unif}(0, 1), \tag{2.7.3}$$

where $\mathcal{V}^-_{z_M}(y)$ *and* $\mathcal{V}^+_{z_M}(y)$ *are defined in* (2.4.2) *and* (2.4.3) *with* $A = A(M, z_M)$ *and* $b = b(M, z_M)$*.*

The derivations of the confidence intervals and hypothesis tests in Section 2.5 remain valid using (2.7.3) as the pivot instead of (2.5.1). Figure 2.10 illustrates the effect of minimal post-selection inference in a simulation study, as compared with the "simple" inference described previously. The intervals are similar in most cases, but one can obtain great gains in precision using the minimal intervals when the simple intervals are very wide.

However, the tradeoff for this increased precision is greater computational cost. We computed $\mathcal{V}^-_{z_M}$ and $\mathcal{V}^+_{z_M}$ for all $z_M \in \{-1, 1\}^{|M|}$, which is only feasible when $|M|$ is fairly small. In what follows, we revert to the simple intervals described in Section 2.5, but extensions to the minimal inference setting are straightforward.

2.8 Extensions

2.8.1 Elastic net

One problem with the lasso is that it tends to select only one variable out of a set of correlated variables, resulting in estimates which are unstable. The elastic net (Zou and Hastie, 2005) adds an ℓ_2 penalty to the lasso objective in order to stabilize the estimates:

$$\hat{\beta}^e = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 + \frac{\gamma}{2} \|\beta\|_2^2 . \tag{2.8.1}$$

Using a nearly identical argument to the one in Section 2.3, we see that necessary and sufficient conditions for $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$ are the existence of $U(M, z_M)$ and $W(M, z_M)$ satisfying

$$\begin{aligned} (X_M^T X_M + \gamma I)U - X_M^T y + \lambda z_M &= 0 \\ X_{-M}^T X_M U - X_{-M}^T y + \lambda W &= 0 \\ \operatorname{sign}(U) = z_M, \;\; W &\in (-1, 1). \end{aligned}$$

Solving for U and W, we see that the selection event can be written

$$\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\} = \left\{ \begin{pmatrix} A_0(M, z_M) \\ A_1(M, z_M) \end{pmatrix} y < \begin{pmatrix} b_0(M, z_M) \\ b_1(M, z_M) \end{pmatrix} \right\} \tag{2.8.2}$$

where A_0, A_1, b_0, and b_1 are the same as in Proposition 2.3.2, except replacing $(X_M^T X_M)^{-1}$, which appears in the expressions through P_M and $(X_M^T)^+$, by the "damped" version $(X_M^T X_M + \gamma I)^{-1}$.

Having rewritten the selection event in the form (2.8.2), we can once again apply the framework of Section 2.4 to obtain a test for the elastic net conditional on this event.

2.8.2 Alternative norms as test statistics

In Section 2.5.2 we used the test statistic

$$T_\infty = \|X_{-\hat{M}}^T (I - P_{\hat{M}}) y\|_\infty$$

and its conditional distribution on $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$ to test whether we had missed any large partial correlations in using $\hat{M}$ as the estimated active set. If we have indeed missed some variables in M there is no reason to suppose that the mean of $X_{-M}^T(I - P_M)y$ is sparse; hence the ℓ_∞ norm may not be the best norm to use as a test statistic.

In principle, we could have used virtually any norm, as long as we can say something about the distribution of this norm conditional on $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$. Problems of this form are considered in Taylor et al. (2013). For example, if we consider the quadratic

$$T_2 = \|X_{-M}^T(I - P_M)y\|_2$$

the general approach in Taylor et al. (2013) derives the conditional distribution of T_2 conditioned on

$$\eta_2^* = \arg\max_{\|\eta\|_2 \leq 1} \eta^T (X_{-M}^T(I - P_M)y).$$

In general, this distribution will be a χ^2 subject to random truncation as in Section 2.4 (see the group lasso examples in Taylor et al. (2013)). Adding the constraints encoded by $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$ affects only the random truncation $[\mathcal{V}^-, \mathcal{V}^+]$.

2.8.3 Estimation of σ^2

As noted above, all of our results rely on a reliable estimate of σ^2. While there are several approaches to estimating σ^2 in the literature, the truncated Gaussian theory described in this work itself provides a natural estimate.

Suppose the linear model is correct ($\mu = X\beta^0$). Then, on the event $\{\hat{M} = M, \hat{E} \supset S\}$, which we assume, the residual

$$(I - P_M)y$$

is a (multivariate) truncated Gaussian with mean $\mathbf{0}$, with law

$$\mathbb{P}_{C,\sigma^2}(B) = \mathbb{P}(Z \in B | Z \in C), \qquad Z \sim N(\mathbf{0}, \sigma^2 I).$$

As σ^2 varies, one obtains a one-parameter exponential family with density

$$\frac{d\mathbb{P}_{C,\sigma^2}}{dz} = e^{-\alpha\|z\|_2^2 - \Lambda_C(\alpha)} 1_C(z)$$

and natural parameter $\alpha = \sigma^2/2$. On the event $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$, we set

$$C = \{y : A(M, z_M)y \leq b(M, z_M)\},$$

and then choose α (or equivalently, σ^2) to satisfy the score equation

$$\mathbb{E}_{C,\hat{\sigma}^2}(\|Z\|_2^2) = \|(I - P_M)y\|_2^2. \tag{2.8.3}$$

This amounts to a maximum likelihood estimate of σ^2. The expectation on the left is generally impossible to do analytically, but there exist fast algorithms for sampling from $\mathbb{P}_{C,\sigma^2}$, c.f. Geweke (1991); Rodriguez-Yam et al. (2004). A rough outline of a naive version of such algorithms is to pick a direction such as e_i one of the coordinate axes. Based on the current state of Z, draw a new entry for the Z_i from the appropriate univariate truncated normal determined from the cutoffs described in Section 2.4. We repeat this procedure to evaluate the expectation on the left, and use gradient descent to find $\hat{\sigma}^2$.

2.8.4 Composite Null Hypotheses

In Section 2.5, we considered hypotheses of the form $H_0 : \eta_{j\star}^T \mu = 0$, which said that the partial correlation of the variables in $-M$ with y, adjusting for the variables in M, was exactly 0. This may be unrealistic, and in practice, we may want to allow some tolerance for the partial correlation.

We consider testing instead the *composite* hypothesis

$$H_0 : |\eta_{j\star}^T \mu| \leq \delta_0. \tag{2.8.4}$$

The following result characterizes a test for H_0.

Proposition 2.8.1. *The test which rejects when* $F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\delta_0,\ \sigma^2||\eta_{j\star}||^2}(\eta^T y) > 1 - \alpha$ *is exact level* α.

Proof. Let $\delta := \eta_{j^\star}^T \mu$. Define $T_{\delta_0} := \inf_{|\delta| \leq \delta_0} F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\delta,\ \sigma^2||\eta_{j^\star}||^2}(\eta^T y)$. Then:

$$\begin{aligned}
\text{Type I error} &:= \sup_{|\delta| \leq \delta_0} \mathbb{P}_\delta(T_{\delta_0} > 1 - \alpha) \\
&\leq \sup_{|\delta| \leq \delta_0} \mathbb{P}_\delta \left(F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\delta,\ \sigma^2||\eta_{j^\star}||^2}(\eta^T y) > 1 - \alpha \right) \\
&= \alpha
\end{aligned}$$

Next, we have that $T_{\delta_0} = F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\delta_0,\ \sigma^2||\eta_{j^\star}||^2}(\eta^T y)$, i.e., the infimum is achieved at $\delta = \delta_0$, so calculating T_{δ_0} is a simple matter of evaluating F_{δ_0}. This follows from the fact that F_δ is monotone decreasing in δ (c.f. Appendix 2.10.1).

Finally, the Type I error is exactly α because the reverse inequality also holds:

$$\text{Type I error} \geq \mathbb{P}_{\delta_0}(T_{\delta_0} > 1 - \alpha) = \alpha.$$

□

Although the test is exact level α, the significance level of a test for a composite null is a "worst-case" Type I error; for most values of μ such that $|\eta^T \mu| \leq \delta_0$, the Type I error will be less than α, so the test will be conservative. Of course, what we lose in power, we gain in robustness to the assumption that $\eta^T \mu = 0$ exactly.

2.8.5 How long a lasso should you use?

Procedures for fitting the lasso, such as `glmnet` (Friedman et al., 2010b), solve (2.1.2) for a decreasing sequence of λ values starting from $\lambda_1 = \|X^T y\|_\infty$. The framework developed so far provides a means to decide when to stop along the regularization path, i.e., when the lasso has done enough "fitting." In this section, we describe a path-wise testing procedure for the lasso,

The path-wise procedure is simple. At each value of λ:

1. Solve the lasso and obtain an active set $\hat{M}_\lambda$ and signs $\hat{z}_{\hat{M}_\lambda}$.
2. Test $H_{0,\lambda} : X^T_{\hat{E}_\lambda}(I - P_{\hat{E}_\lambda})(\mu) = 0$ at level α. Rather than being conditional on only $(\hat{M}_\lambda, \hat{z}_{\hat{M}_\lambda})$, this test is conditional on the entire sequence of active sets and signs $\{(\hat{M}^m, \hat{z}^m) = (M^m, z^m)\}$, as we describe below.

As λ decreases, we expect to reject the null hypotheses as the fit improves and stop once the first null hypothesis has been accepted.

To understand the properties of this procedure, we formalize it as a multiple testing problem. For each value $\lambda_1, ..., \lambda_m$, we test H_{0,λ_i}. We test these hypotheses sequentially and stop after the first hypothesis has been accepted. Implicitly, this means that we accept all the remaining hypotheses.

Our next result shows that this procedure controls the family-wise error rate (FWER) at level α. Let V denote that number of false rejections. Then FWER is defined as $\mathbb{P}(V \geq 1)$. The practical implication of this result is the model selected by this procedure will be larger than the true model with probability α.

Proposition 2.8.2. *The path-wise testing procedure controls FWER at level α.*

Proof. Let $\hat{M}^m$ and $\hat{z}^m$ denote the complete sequence of active sets and signs at $\lambda_1, \ldots, \lambda_m$, i.e.,

$$\hat{M}^m = \{\hat{M}_{\lambda_1}, \ldots, \hat{M}_{\lambda_m}\}$$
$$\hat{z}^m = \{\hat{z}_{\hat{M}_{\lambda_1}}, \ldots, \hat{z}_{\hat{M}_{\lambda_m}}\}.$$

We seek to control the family-wise error rate (FWER) when testing the hypotheses $H_{0,\lambda_1}, \ldots, H_{0,\lambda_m}$, i.e., $\mathbb{P}(V \geq 1)$. We partition the space over all possible sequences $\hat{M}^m$ and $\hat{z}^m$:

$$\mathbb{P}(V \geq 1) = \sum_{(M^m, z^m)} \mathbb{P}\left(V \geq 1 \mid (\hat{M}^m, \hat{z}^m) = (M^m, z^m)\right) \mathbb{P}\left((\hat{M}^m, \hat{z}^m) = (M^m, z^m)\right).$$

Since $\sum_{(M^m, z^m)} \mathbb{P}\left((\hat{M}^m, \hat{z}^m) = (M^m, z^m)\right) = 1$, we can ensure FWER $\leq \alpha$ by ensuring

$$\mathbb{P}\left(V \geq 1 \mid (\hat{M}^m, \hat{z}^m) = (M^m, z^m)\right) \leq \alpha \text{ for any } (M^m, z^m).$$

Let λ_k denote the first λ_i for which H_{0,λ_i} is true. Then the event $V \geq 1$ is equivalent to the event that we reject H_{0,λ_k} because the preceding hypotheses $H_{0,\lambda_1}, \ldots, H_{0,\lambda_{k-1}}$ are all false so we cannot make a false discovery before the k^{th} hypothesis. Thus

$$\mathbb{P}\left(V \geq 1 \mid (\hat{M}^m, \hat{z}^m) = (M^m, z^m)\right) = \mathbb{P}\left(\text{reject } H_{0,\lambda_k} \mid (\hat{M}^m, \hat{z}^m) = (M^m, z^m)\right).$$

Therefore, we can control FWER at level α by ensuring

$$\mathbb{P}\left(\text{reject } H_{0,\lambda} \mid (\hat{M}^m, \hat{z}^m) = (M^m, z^m)\right) \leq \alpha$$

for each $\lambda \in \{\lambda_1, \ldots, \lambda_k\}$. □

To perform a test of $H_{0,\lambda}$ conditioned on $\{(\hat{M}^m, \hat{z}^m) = (M^m, z^m)\}$, we apply the framework of Section 2.4. Let

$$\{A(M_i, s_i)y < b(M_i, s_i)\}$$

be the affine constraints that characterize the event $\{(\hat{M}_{\lambda_i}, \hat{z}_{\lambda_i}) = (M_i, z_i)\}$ from Proposition 2.3.2. The event $\{(\hat{M}^m, \hat{z}^m) = (M^m, z^m)\}$ is equivalent to the intersection of all of these constraints:

$$\underbrace{\begin{bmatrix} A(M_1, z_1) \\ \vdots \\ A(M_m, z_m) \end{bmatrix}}_{A(M^m, z^m)} y < \underbrace{\begin{bmatrix} b(M_1, z_1) \\ \vdots \\ b(M_m, z_m) \end{bmatrix}}_{b(M^m, z^m)}.$$

Now Theorem 2.4.2 applies, and we can obtain the usual pivot as a test statistic.

2.9 Conclusion

We have described a method for making inference about $\eta^T\mu$ in the linear model based on the lasso estimator, where η is chosen adaptively after model selection. The confidence intervals and tests that we propose are conditional on $\{(\hat{M}, \hat{z}_{\hat{M}}) = (M, z_M)\}$. In contrast to existing procedures on inference for the lasso, we provide a pivot whose conditional distribution can be characterized exactly (non-asymptotically). This pivot can be used to derive confidence intervals and hypothesis tests based on lasso estimates anywhere along the solution path, not necessarily just at the knots of the LARS path as in Lockhart et al. (2014). Finally, our test is computationally simple: the quantities required to form the test statistic are readily available from the solution of the lasso.

2.10 Appendix

2.10.1 Monotonicity of F

Lemma 2.10.1. *Let $F_\mu(x) := F^{[a,b]}_{\mu,\sigma^2}(x)$ denote the cumulative distribution function of a truncated Gaussian random variable, as defined as in (2.4.5). Then $F_\mu(x)$ is monotone decreasing in μ.*

Proof. First, the truncated Gaussian distribution with CDF $F_\mu := F^{[a,b]}_{\mu,\sigma^2}$ is a natural exponential family in μ, since it is just a Gaussian with a different base measure. Therefore, it has monotone likelihood ratio in μ. That is, for all $\mu_1 > \mu_0$ and $x_1 > x_0$:

$$\frac{f_{\mu_1}(x_1)}{f_{\mu_0}(x_1)} > \frac{f_{\mu_1}(x_0)}{f_{\mu_0}(x_0)}$$

where $f_{\mu_i} := dF_{\mu_i}$ denotes the density. (Instead of appealing to properties of exponential families, this property can also be directly verified.)

This implies

$$f_{\mu_1}(x_1)f_{\mu_0}(x_0) > f_{\mu_1}(x_0)f_{\mu_0}(x_1) \qquad\qquad x_1 > x_0.$$

Therefore, the inequality is preserved if we integrate both sides with respect to x_0 on $(-\infty, x)$ for $x < x_1$. This yields:

$$\int_{-\infty}^{x} f_{\mu_1}(x_1)f_{\mu_0}(x_0)\,dx_0 > \int_{-\infty}^{x} f_{\mu_1}(x_0)f_{\mu_0}(x_1)\,dx_0 \qquad x < x_1$$

$$f_{\mu_1}(x_1)F_{\mu_0}(x) > f_{\mu_0}(x_1)F_{\mu_1}(x) \qquad x < x_1$$

Now we integrate both sides with respect to x_1 on (x, ∞) to obtain:

$$(1 - F_{\mu_1}(x))F_{\mu_0}(x) > (1 - F_{\mu_0}(x))F_{\mu_1}(x)$$

which establishes $F_{\mu_0}(x) > F_{\mu_1}(x)$ for all $\mu_1 > \mu_0$. □

2.11 Lasso Screening Property

In this section, we state some sufficient conditions that guarantee $\text{support}(\beta^0) \subset \text{support}(\hat{\beta})$. Let $M = \text{support}(\beta^0)$ and $\hat{M} \subset \text{support}(\hat{\beta})$. The results of this section are well known in

the literature and can be found in (Bühlmann and van de Geer, 2011, Chapter 2.5).

Definition 2.11.1 (Restricted Eigenvalue Condition)**.** *Restricted eigenvalue condition requires that X satisfy*

$$\|Xv\|_2^2 \geq m \|v\|_2^2$$

for all $v \in \{x : \|x_{-M}\|_1 \leq 3 \|x_M\|\}$.

Definition 2.11.2 (Beta-min Condition)**.** *The beta-min condition requires that for all $j \in M$,*

$$|\beta_j^0| > \beta_{min}.$$

Theorem 2.11.3. *Let $y = X\beta^0 + \epsilon$, where ϵ is subgaussian with parameter σ, and $\hat{\beta}$ be the solution to 2.1.2 with $\lambda = 4\sigma\sqrt{\frac{\log p}{n}}$. Assume that X satisfies the restricted eigenvalue condition, β^0 satisfies the beta-min condition with $\beta_{min} = \frac{8\sigma}{m}\sqrt{\frac{s \log p}{n}}$, and X is column normalized, $\|x_j\|_2 \leq \sqrt{n}$. Then $M \subset \hat{M}$.*

Proof. From (Negahban et al., 2012, Corollary 2),

$$\left\|\hat{\beta} - \beta^0\right\|_2 \leq \frac{8\sigma}{m}\sqrt{\frac{s \log p}{n}}.$$

Assume that their is a j such that $j \in M$, but $j \notin \hat{M}$. We must have

$$\left\|\hat{\beta} - \beta^0\right\|_2 > |\beta_j^0| \geq \beta_{min} = \frac{8\sigma}{m}\sqrt{\frac{s \log p}{n}}.$$

This is a contradiction, so for all $j \in M$ we have $j \in \hat{M}$. □

Lemma 2.11.4. *The conditioning set can be rewritten in terms of $\eta^T y$ as follows:*

$$\{Ay \leq b\} = \{\mathcal{V}^-(y) \leq \eta^T y \leq \mathcal{V}^+(y), \mathcal{V}^0(y) \geq 0\}$$

where

$$\alpha = \frac{A\Sigma\eta}{\eta^T\Sigma\eta} \tag{2.11.1}$$

$$\mathcal{V}^- = \mathcal{V}^-(y) = \max_{j:\ \alpha_j<0} \frac{b_j - (Ay)_j + \alpha_j\eta^T y}{\alpha_j} \tag{2.11.2}$$

$$\mathcal{V}^+ = \mathcal{V}^+(y) = \min_{j:\ \alpha_j>0} \frac{b_j - (Ay)_j + \alpha_j\eta^T y}{\alpha_j}. \tag{2.11.3}$$

$$\mathcal{V}^0 = \mathcal{V}^0(y) = \min_{j:\ \alpha_j=0} b_j - (Ay)_j \tag{2.11.4}$$

Moreover, $(\mathcal{V}^+, \mathcal{V}^-, \mathcal{V}^0)$ *are independent of* $\eta^T y$.

Proof. Although the proof of Lemma 2.11.4 is elementary, the geometric picture gives more intuition as to why $\mathcal{V}^+$ and $\mathcal{V}^-$ are independent of $\eta^T y$. Since Σ is assumed known, let $\tilde{y} = \Sigma^{-\frac{1}{2}} y$ so that $\tilde{y} \sim N(\Sigma^{-\frac{1}{2}}\mu, I)$. We can decompose $\tilde{y}$ into two independent components: a one-dimensional component along $\tilde{\eta} := \Sigma^{\frac{1}{2}}\eta$ and a $(p-1)$-dimensional component orthogonal to $\tilde{\eta}$:

$$\tilde{y} = \tilde{y}_{\tilde{\eta}} + \tilde{y}_{\tilde{\eta}^\perp}.$$

From Figure 2.1, it is clear that the extent of the set $\{Ay \le b\} = \{A\Sigma^{\frac{1}{2}}\tilde{y} \le b\}$ (i.e., $\mathcal{V}^+$ and $\mathcal{V}^-$) along the direction $\tilde{\eta}$ depends only on $\tilde{y}_{\tilde{\eta}^\perp}$ and is hence independent of $\tilde{\eta}^T\tilde{y} = \eta^T y$. We present a geometric derivation below. The values $\mathcal{V}^+$ and $\mathcal{V}^-$ are the maximum and minimum possible values of $\tilde{\eta}^T\tilde{y}$, holding $\tilde{y}_{\tilde{\eta}^\perp}$ fixed, while remaining inside the polytope $A\Sigma^{\frac{1}{2}}\tilde{y} \le b$. Writing $\tilde{y} = c\tilde{\eta} + \tilde{y}_{\tilde{\eta}^\perp}$ where c is allowed to vary, $\mathcal{V}^+$ and $\mathcal{V}^-$ are the optimal values of the optimization problems:

$$\begin{aligned} \text{max. / min.} \quad & \tilde{\eta}^T\tilde{y} = c||\tilde{\eta}||_2^2 \\ \text{subject to} \quad & A\Sigma^{\frac{1}{2}}(c\tilde{\eta} + \tilde{y}_{\tilde{\eta}^\perp}) \le b \end{aligned}$$

Rewriting this problem in terms of the original variables η and y, we obtain:

$$\begin{aligned} \text{max. / min.} \quad & c(\eta^T\Sigma\eta) \\ \text{subject to} \quad & c(A\Sigma\eta) \le b - Ay + \frac{A\Sigma\eta}{\eta^T\Sigma\eta}\eta^T y \end{aligned}$$

Since c is the only free variable, we see from the constraints that the optimal values $\mathcal{V}^+$ and $\mathcal{V}^-$ are precisely those given in (2.4.2) and (2.4.3). □

Chapter 3

Condition-on-Selection Method

In the previous chapter, we focused on selective inference for the sub-model coefficients selected by the lasso by conditioning on the event that lasso selects a certain subset of variables. However the procedure we developed is not restricted to the sub-model coefficients, nor is it restricted to the lasso. In Lee and Taylor (2014), we used the same Condition-on-Selection (COS) method for marginal screening, orthogonal matching pursuit, and screening+lasso variable selection methods.

In this chapter, we first discuss some definitions and formalism, which will help us understand how to generalize the results of Chapter 2 to other selection procedures. In Section 3.1, we see that the COS method results in tests that control the selective type 1 error. Then in Section 3.2, we show how the selection events for several variable selection methods such as marginal screening, and orthogonal matching pursuit are affine in the response y. For non-affine selection events, we propose a general algorithm in Section 3.3. We then describe inference for the full model regression coefficients, provide a method for FDR control and establish the asymptotic coverage property in the high-dimensional setting in Section 3.4. Finally in Section 3.5, we show how to construct selectively valid confidence intervals for regression coefficients selected by the knockoff filter.

3.1 Formalism

This section closely follows the development in Fithian et al. (2014), which in turn uses the COS method developed in earlier works Lee and Taylor (2014); Lee et al. (2013a); Taylor et al. (2014). Our main result of this section is to show that tests constructed using the

COS method control selective type 1 error. This is the original motivation of Lee and Taylor (2014); Lee et al. (2013a) for designing tests with the COS method.

We start off by defining a valid test in the classical setting.

Definition 3.1.1 (Valid test)**.** *Let $H \in \mathcal{H}$ be a hypothesis, and $\phi(y; H) \in \{0, 1\}$ is a test of H meaning we reject H if $\phi(y; H) = 1$. $\phi(y; H)$ is a valid test of H if*

$$\mathbb{P}_F\left(\phi(y; H) = 1\right) \leq \alpha$$

for all F null with respect to H, meaning $F \in N_H$, where N_H is the set of distributions null with respect to H.

For selective inference, there is an analog of type 1 error.

Definition 3.1.2 (Selective Type 1 Error)**.** *$\phi(y, H(y))$ is a valid test of the hypothesis $H(y)$ if it controls the selective type 1 error,*

$$\mathbb{P}_F\left(\phi(y; H(y)) = 1 \mid F \in N_{H(y)}\right) \leq \alpha.$$

The framework laid out in Chapter 2 proposes controlling the selective type 1 error via the COS method. As we showed in the case of confidence intervals for regression coefficients and goodness-of-fit tests, by conditioning on the lasso selection event, we are guaranteed to control the conditional type 1 error by design, and this implies the control of the unconditional type 1 error. We now show that this is not specific to the lasso; in fact controlling the conditional type 1 error always controls the unconditional type 1 error in Definition 3.1.2.

Definition 3.1.3. *Let $\mathcal{H}$ be the hypothesis space. The selection algorithm $H : \mathbf{R}^n \to \mathcal{A}$ maps data to hypothesis. This induces the selection event $S(H) = \{y : H(y) = H\}$.*

The following definition motivates the construction in Equation (2.5.3).

Definition 3.1.4 (Condition-on-Selection method)**.** *A test ϕ is constructed via the Condition-on-Selection (COS) method if for all $F \in N_{H_i}$*

$$\mathbb{P}_F\left(\phi(y; H_i) = 1 \mid H(y) = H_i\right). \tag{3.1.1}$$

This means that $\phi(y; H_i)$ controls the conditional type 1 error rate.

By a simple generalization of the argument in Theorem 2.4.3, we show that using the COS method to design a conditional test 3.1.1 implies control of the selective type 1 error 3.1.2.

Theorem 3.1.5 (Selective Type 1 Error control)**.** *A test constructed using the COS method, i.e. satisfies* (3.1.1)*, controls the selective type 1 error meaning*

$$\mathbb{P}_F\left(\phi(y; H(y)) = 1 \mid F \in N_{H(y)}\right) \leq \alpha.$$

Proof.

$$\begin{aligned}
&\mathbb{P}_F(\phi(y; H(y)) = 1 \mid F \in N_{H(y)}) = \sum_{i=1}^{|\mathcal{H}|} \mathbb{P}(\phi(y; H(y)) = 1, H(y) = H_i) \mid F \in N_{H(y)}) \\
&= \sum_{i:F\in N_{H_i}} \mathbb{P}(\phi(y; H(y)) = 1, H(y) = H_i \mid F \in N_{H(y)}) + \\
&\sum_{i:F\notin N_{H_i}} \mathbb{P}(\phi(y; H(y)) = 1, H(y) = H_i \mid F \in N_{H(y)}) \\
&= \sum_{i:F\in N_{H_i}} \mathbb{P}(\phi(y; H(y)) = 1, H(y) = H_i \mid F \in N_{H(y)}) + 0 \\
&= \sum_{i:F\in N_{H_i}} \mathbb{P}(\phi(y; H(y)) = 1 \mid F \in N_{H(y)}, H(y) = H_i)\mathbb{P}(H(y) = H_i \mid F \in N_{H(y)}) \\
&= \sum_{i:F\in N_{H_i}} \mathbb{P}(\phi(y; H_i) = 1 \mid H(y) = H_i)\mathbb{P}(H(y) = H_i \mid F \in N_{H(y)}) \\
&= \sum_{i:F\in N_{H_i}} \alpha\mathbb{P}(H(y) = H_i \mid F \in N_{H(y)}) \\
&\leq \alpha.
\end{aligned}$$

where all of the previous probabilities are with respect to the distribution F. The first equality is the law of total probability, and the second equality is breaking the sum over disjoint sets. Since $F \notin N_{H_i}$, implies $\mathbb{P}_F(H(y) = H_i \mid F \in N_{H(y)}) = 0$, so $\sum_{i:F\notin H_i} \mathbb{P}(\phi(y; H(y)) = 1, H(y) = H_i \mid F \in N_{H(y)}) = 0$, which establishes the third equality. The fourth equality is the definition of conditional probability, and the fifth follows from noticing that $\{F \in N_{H(y)}, H(y) = H_i, F \in N_{H_i}\} = \{H(y) = H_i, F \in N_{H_i}\}$. The sixth equality uses the

COS property of ϕ: $\mathbb{P}_F(\phi(y; H_i) = 1 \mid H(y) = H_i) \leq \alpha$ for any $F \in N_{H_i}$. Finally, the result follows since probabilities sum to less than or equal to 1. □

This result allows us to interpret the tests constructed via the COS method as unconditionally valid.

3.2 Marginal Screening, Orthogonal Matching Pursuit, and other Variable Selection methods

In lieu of the developments of the previous section, it is clear that the COS method developed for affine selection events in Chapter 2 is not specific to the lasso. By changing the variable selection method, we are simply changing the selection algorithm and the selection event. The main work is in characterizing the selection event $\{y : \hat{M}(y) = M\}$, the event that the variable selection methods chooses the subset M. In this section, we characterize the selection event for several variable selection methods: marginal screening, orthogonal matching pursuit (forward stepwise), non-negative least squares, and marginal screening+lasso.

3.2.1 Marginal Screening

In the case of marginal screening, the selection event $\hat{M}(y)$ corresponds to the set of selected variables $\hat{M}$ and signs s:

$$\begin{aligned}\hat{M}(y) &= \left\{y : \text{sign}(x_i^T y) x_i^T y > \pm x_j^T y \text{ for all } i \in \hat{M} \text{ and } j \in \hat{M}^c\right\} \\ &= \left\{y : \hat{s}_i x_i^T y > \pm x_j^T y \text{ and } \hat{s}_i x_i^T y \geq 0 \text{ for all } i \in \hat{M} \text{ and } j \in \hat{M}^c\right\} \\ &= \left\{y : A(\hat{M}, \hat{s}) y \leq 0\right\} \end{aligned} \tag{3.2.1}$$

for some matrix $A(\hat{M}, \hat{s})$.

3.2.2 Marginal screening + Lasso

The marginal screening+Lasso procedure was introduced in Fan and Lv (2008) as a variable selection method for the ultra-high dimensional setting of $p = O(e^{n^k})$. Fan et al. Fan and Lv (2008) recommend applying the marginal screening algorithm with $k = n - 1$, followed by the Lasso on the selected variables. This is a two-stage procedure, so to properly account

for the selection we must encode the selection event of marginal screening followed by Lasso. This can be done by representing the two stage selection as a single event. Let $(\hat{M}_m, \hat{s}_m)$ be the variables and signs selected by marginal screening, and the $(\hat{M}_L, \hat{z}_L)$ be the variables and signs selected by Lasso. In Proposition 2.2 of Lee et al. (2013a), it is shown how to encode the Lasso selection event $(\hat{M}_L, \hat{z}_L)$ as a set of constraints $\{A_L y \leq b_L\}$ [1], and in Section 3.2.1 we showed how to encode the marginal screening selection event $(\hat{M}_m, \hat{s}_m)$ as a set of constraints $\{A_m y \leq b_m\}$. Thus the selection event of marginal screening+Lasso can be encoded as $\{A_L y \leq b_L, A_m y \leq 0\}$.

3.2.3 Orthogonal Matching Pursuit

Orthogonal matching pursuit (OMP) is a commonly used variable selection method [2]. At each iteration, OMP selects the variable most correlated with the residual r, and then recomputes the residual using the residual of least squares using the selected variables. The description of the OMP algorithm is given in Algorithm 1.

Algorithm 1 Orthogonal matching pursuit (OMP)

1: **Input:** Design matrix X, response y, and model size k.
2: **for**: $i = 1$ to k
3: $\quad p_i = \arg\max_{j=1,\ldots,p} |r_i^T x_j|$.
4: $\quad \hat{S}_i = \cup_{j=1}^{i} \{p_i\}$.
5: $\quad r_{i+1} = (I - X_{\hat{S}_i} X_{\hat{S}_i}^+) y$.
6: **end for**
7: **Output**: $\hat{S} := \{p_1, \ldots, p_k\}$, and $\hat{\beta}_{\hat{S}} = (X_{\hat{S}}^T X_{\hat{S}})^{-1} X_{\hat{S}}^T y$

The OMP selection event as a set of linear constraints on y.

$$\begin{aligned}
\hat{M}(y) &= \left\{ y : \text{sign}(x_{p_i}^T r_i) x_{p_i}^T r_i > \pm x_j^T r_i, \text{ for all } j \neq p_i \text{ and all } i \in [k] \right\} \\
&= \{ y : \hat{s}_i x_{p_i}^T (I - X_{\hat{M}_{i-1}} X_{\hat{M}_{i-1}}^+) y > \pm x_j^T (I - X_{\hat{M}_{i-1}} X_{\hat{M}_{i-1}}^+) y \text{ and} \\
&\quad \hat{s}_i x_{p_i}^T (I - X_{\hat{M}_{i-1}} X_{\hat{M}_{i-1}}^+) y > 0, \text{ for all } j \neq p_i, \text{ and all } i \in [k] \ \} \\
&= \left\{ y : A(\hat{M}_1, \ldots, \hat{M}_k, \hat{s}_1, \ldots, \hat{s}_k) \leq b(\hat{M}_1, \ldots, \hat{M}_k, \hat{s}_1, \ldots, \hat{s}_k) \right\}.
\end{aligned}$$

[1] The Lasso selection event is with respect to the Lasso optimization problem after marginal screening.
[2] OMP is sometimes known as forward stepwise regression.

The selection event encodes that OMP selected a certain variable and the sign of the correlation of that variable with the residual, at steps 1 to k. The primary difference between the OMP selection event and the marginal screening selection event is that the OMP event also describes the order at which the variables were chosen. The marginal screening event only describes that the variable was among the top k most correlated, and not whether a variable was the most correlated or kth most correlated.

3.2.4 Nonnegative Least Squares

Non-negative least squares (NNLS) is a simple modification of the linear regression estimator with non-negative constraints on β:

$$\arg\min_{\beta:\beta\geq 0} \frac{1}{2}\left\|y - X\beta\right\|^2. \tag{3.2.2}$$

Under a positive eigenvalue conditions on X, several authors Slawski et al. (2013); Meinshausen et al. (2013) have shown that NNLS is comprable to the Lasso in terms of prediction and estimation errors. The NNLS estimator also does not have any tuning parameters, since the sign constraint provides a natural form of regularization. NNLS has found applications when modeling non-negative data such as prices, incomes, count data. Non-negativity constraints arise naturally in non-negative matrix factorization, signal deconvolution, spectral analysis, and network tomography; we refer to Chen and Plemmons (2009) for a comprehensive survey of the applications of NNLS.

We show how our framework can be used to form exact hypothesis tests and confidence intervals for NNLS estimated coefficients. The primal dual solution pair $(\hat{\beta}, \hat{\lambda})$ is a solution iff the KKT conditions are satisfied,

$$\hat{\lambda}_i := -x_i^T(y - X\hat{\beta}) \geq 0 \text{ for all i}$$
$$\hat{\beta} \geq 0.$$

Let $\hat{M} = \{i : -x_i^T(y - X\hat{\beta}) = 0\}$. By complementary slackness $\hat{\beta}_{-\hat{M}} = 0$, where $-\hat{M}$ is the complement to the "active" variables $\hat{M}$ chosen by NNLS. Given the active set we can

solve the KKT equation for the value of $\hat{\beta}_{\hat{M}}$,

$$\begin{aligned} -X_{\hat{M}}^T(y - X\hat{\beta}) &= 0 \\ -X_{\hat{M}}^T(y - X_{\hat{M}}\hat{\beta}_{\hat{M}}) &= 0 \\ \hat{\beta}_{\hat{M}} &= X_{\hat{M}}^+ y, \end{aligned}$$

which is a linear contrast of y. The NNLS selection event is

$$\begin{aligned} \hat{M}(y) &= \{y : X_{\hat{M}}^T(y - X\hat{\beta}) = 0,\ X_{-\hat{M}}^T(y - X\hat{\beta}) > 0\} \\ &= \{y : X_{\hat{M}}^T(y - X\hat{\beta}) \geq 0, -X_{\hat{M}}^T(y - X\hat{\beta}) \geq 0, X_{-\hat{M}}^T(y - X\hat{\beta}) > 0\} \\ &= \{y : X_{\hat{M}}^T(I - X_{\hat{M}}X_{\hat{M}}^+)y \geq 0, -X_{\hat{M}}^T(I - X_{\hat{M}}X_{\hat{M}}^+)y \geq 0, X_{-\hat{M}}^T(I - X_{\hat{M}}X_{\hat{M}}^+)y > 0\} \\ &= \{y : A(\hat{M})y \leq 0\}. \end{aligned}$$

The selection event encodes that for a given y the NNLS optimization program will select a subset of variables $\hat{M}(y)$.

3.2.5 Logistic regression with Screening

The focus up to now has been on the linear regression estimator with additive Gaussian noise. In this section, we discuss extensions to conditional MLE (maximum likelihood estimator) such as logistic regression. This section is meant to be speculative and non-rigorous; our goal is only to illustrate that these tools are not restricted to the linear regression. A future publication will rigorously develop the inferential framework for conditional MLE.

Consider the logistic regression model with loss function and gradient below,

$$\begin{aligned} \ell(\beta) &= \frac{1}{n}\left(-y^T X\beta + \sum_{i=1}^n \log(1 + e^{\beta^T x_i})\right) \\ \nabla\ell(\beta) &= -\frac{1}{n}X^T(y - s(X\beta)), \end{aligned}$$

where $s(X\beta)$ is the sigmoid function applied entrywise. By taylor expansion, the empirical

estimator is given by

$$\begin{aligned}\hat{\beta} &\approx \beta^0 - \left(\nabla^2 \ell(\beta^0)\right)^{-1} \nabla \ell(\beta^0) \\ &= \beta^0 + \left(\nabla^2 \ell(\beta^0)\right)^{-1} X^T (y - s(X\beta))\end{aligned}$$

By the Lindeberg CLT (central limit theorem), $\frac{1}{\sqrt{n}} X^T (y - s(X\beta^0)) \to \mathcal{N}(0, \mathbf{E}(\nabla^2 \ell(\beta^0)))$, and thus $w := \frac{1}{\sqrt{n}} X^T y$ converges to a Gaussian. The marginal screening selection procedure can be expressed as a set of inequalities $\{\text{sign}(w_i) w_i \geq \pm w_j, i \in \hat{M}, j \in \hat{M}^c\} = \{Aw \leq b\}$. Thus conditional on the selection, w is approximately a constrained Gaussian. The framework in Chapter 2.4 and 3.1 can be applied to w, instead of y, to derive hypothesis tests and confidence intervals for the coefficients of logistic regression. The resulting test and confidence intervals should be correct asymptotically. However, this is the best we can expect for logistic regression and other conditional MLE because even in the classical case the Wald test is only asymptotically correct. For other conditional maximum likelihood estimator similar reasoning applies, since the gradient $\nabla \ell(\beta)$ converges in distribution to a Gaussian.

For logistic regression with ℓ_1 regularizer,

$$\frac{1}{n} \left(-y^T X\beta + \sum_{i=1}^{n} \log(1 + e^{\beta^T x_i}) \right) + \lambda \|\beta\|_1$$

the selection event cannot be analytically described. However, the COS method can still be applied using the general method presented in Chapter 3.3.

3.3 General method for Selective inference

In this section, we describe a computationally-intensive algorithm for finding selection events, when they are not easily described analytically.

We first review the construction used in Chapter 2 for affine selection events. Let $P^{\perp}_{\Sigma,\eta}(y) = (I - \frac{\Sigma\eta\eta^T}{\eta^T\Sigma\eta})y$. Recall that y can be decomposed into two independent components $y = (\eta^T y)\frac{\Sigma\eta}{\eta^T\Sigma\eta} + (I - \frac{\Sigma\eta\eta^T}{\eta^T\Sigma\eta})y$. This is derived by defining $\tilde{y} = \Sigma^{-1/2} y \sim \mathcal{N}(0, I)$ and

$\tilde{\eta} = \Sigma^{1/2}y$. $\tilde{y}$ can be orthogonally decomposed as $\tilde{y} = (\tilde{\eta}^T\tilde{y})\frac{\tilde{\eta}}{\|\tilde{\eta}\|} + (I - \frac{\tilde{\eta}\tilde{\eta}^T}{\|\tilde{\eta}\|^2})\tilde{y}$, so

$$\begin{aligned} y = \Sigma^{1/2}\tilde{y} &= (\tilde{\eta}^T\tilde{y})\frac{\Sigma^{1/2}\tilde{\eta}}{\|\tilde{\eta}\|} + \Sigma^{1/2}(I - \frac{\tilde{\eta}\tilde{\eta}^T}{\|\tilde{\eta}\|^2})\tilde{y} \\ &= (\eta^T y)\frac{\Sigma\eta}{\eta^T\Sigma\eta} + (I - \frac{\Sigma\eta\eta^T}{\eta^T\Sigma\eta})y. \end{aligned}$$

Lemma 2.4.1 shows that

$$\eta^T y|\{Ay \leq b, P^{\perp}_{\Sigma,\eta}y = y_0\} \sim TN(\eta^T\mu, \sigma^2\|\eta\|^2, \mathcal{V}^-(y_0, A, b), \mathcal{V}^+(y_0, A, b)).$$

We can generalize this result to arbitrary selection events, where the selection event is not explicitly describable. Recall that H is a selection algorithm that maps $\mathbf{R}^n \to \mathcal{H}$. The selection event is $S(H) = \{x : H(x) = H\}$, so $y \in S(H)$ iff $H(y) = H$. In the upcoming section, it will be convenient to work with the definition using $H(\cdot)$, since the set $S(H)$ cannot be described, but the function $H(\cdot)$ can be efficiently computed. Thus we can only verify if a point $y \in S(H)$.

For testing slices $\eta^T\mu$, we only need the distribution of $\eta^T y|\{S, P^{\perp}_{\Sigma,\eta}(y) = y_0\}$.

The following Theorem is a straightforward generalization of Theorem 2.4.2 from polyhedral sets to arbitrary sets S.

Theorem 3.3.1 (Arbitrary selection events)**.** *Let y be a multivariate truncated normal, so $L(y) \propto e(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu))\mathbf{1}(y \in S(H))$. Then*

$$\eta^T y|\{y \in S(H), P^{\perp}_{\Sigma,\eta}y = y_0\} \stackrel{d}{=} TN(\eta^T\mu, \eta^T\Sigma\eta, U(H, y_0, \frac{\Sigma\eta}{\eta^T\Sigma\eta}))$$

and $U(H, y_0, \frac{\Sigma\eta}{\eta^T\Sigma\eta})) = \{c : H(y_0 + c\frac{\Sigma\eta}{\eta^T\Sigma\eta}) = H\}$.

Proof. We know that $\eta^T y|\{y \in S(H), P^{\perp}_{\Sigma,\eta}y = y_0\} \stackrel{d}{=} TN(\eta^T\mu, \|\eta\|^2, U(H, y_0, \frac{\Sigma\eta}{\eta^T\Sigma\eta}))$, so $\eta^T y|\{y \in S(H), P^{\perp}_{\Sigma,\eta}y = y_0\}$ is a univariate normal truncated to some region U. The goal is to check that $U(H, y_0, \Sigma\eta) = \{c : H(y_0 + c\frac{\Sigma\eta}{\eta^T\Sigma\eta}) = H\}$. We can describe the conditioning

set as

$$\begin{aligned}
&\{y : y \in S(H), P^{\perp}_{\Sigma,\eta} y = y_0\} = \{y : H(y) = H, P^{\perp}_{\Sigma,\eta} y = y_0\} \\
&= \{y = y_0 + c\frac{\Sigma\eta}{\eta^T\Sigma\eta} : H(y_0 + c\frac{\Sigma\eta}{\eta^T\Sigma\eta}) = H, P^{\perp}_{\Sigma,\eta} y = y_0\} \\
&= \{y = y_0 + c\frac{\Sigma\eta}{\eta^T\Sigma\eta} : P^{\perp}_{\Sigma,\eta} y = y_0, c \in U(H, y_0, c\frac{\Sigma\eta}{\eta^T\Sigma\eta})\} \\
&= \{y : P^{\perp}_{\Sigma,\eta} y = y_0, \eta^T y \in U(H, y_0, c\frac{\Sigma\eta}{\eta^T\Sigma\eta})\}
\end{aligned}$$

Thus we have that

$$\begin{aligned}
\left[\eta^T y | \{y \in S(H), P^{\perp}_{\Sigma,\eta} y = y_0\}\right] &\overset{d}{=} \left[\eta^T y | \eta^T y \in U(H, y_0, \frac{\Sigma\eta}{\eta^T\Sigma\eta}), P^{\perp}_{\Sigma,\eta} y = y_0\}\right] \\
&\overset{d}{=} \left[\eta^T y | \eta^T y \in U(H, y_0, \frac{\Sigma\eta}{\eta^T\Sigma\eta})\}\right] \\
&\sim TN(\eta^T\mu, \eta^T\Sigma\eta, U(H, y_0, \frac{\Sigma\eta}{\eta^T\Sigma\eta}))
\end{aligned}$$

where the second equality follows from independence of $\eta^T y$ and $P^{\perp}_{\Sigma,\eta} y$. □

3.3.1 Computational Algorithm for arbitrary selection algorithms

In this section, we study the case of where the set $S(H)$ cannot be explicitly described, but the function $H(\cdot)$ is easily computable. Our goal will be to approximately compute the p-value $F(\eta^T y; \eta^T\mu, U(H, y_0, \frac{\Sigma\eta}{\eta^T\Sigma\eta}))$, where F is the cdf of $TN(\eta^T\mu, \eta^T\Sigma\eta, U(H, y_0, \frac{\Sigma\eta}{\eta^T\Sigma\eta})$.

Algorithm 2 is the primary contribution of this section. This allows us to compute the pivotal quantity for algorithms $H(\cdot)$ with difficult to describe selection events. This includes linear regression with the SCAD/MCP regularizers, and logistic regression with ℓ_1-regularizer, where the selection events do not have analytical forms.

Let $\tilde{\phi}(z; \nu, \sigma) = \phi(\frac{z-\nu}{\sigma})$ be the pdf of a univariate truncated normal with mean ν and variance σ^2. Algorithm 2 gives an approximate p-value for the null hypothesis $H_0 : \eta^T\mu = \gamma$. The advantage of this algorithm is it does not need an explicit description of the set S, nor the set U. It runs the selection algorithm $H(\cdot)$ at the grid points d_i, and determines if the point $y_0 + d\frac{\Sigma\eta}{\eta^T\Sigma\eta}$ is in the selection event. Then it approximates the CDF of the univariate truncated normal by a discrete truncated normal.

Conjecture 3.3.2. *Let D_m be a set of grid points $2m^2$ grid points that is equispaced on*

Algorithm 2 Compute approximate p-value

Input: Grid points $D = \{d_1, \ldots, d_n\}$ and empty set $C = \emptyset$
Output: Approximate p-value p
for all $d_i \in D$ **do**
  Compute $H_i = H(y_0 + d\frac{\Sigma\eta}{\eta^T\Sigma\eta})$.
  if $H_i = H$ **then**
    $C = C \cup d_i$.
  end if
end for
Return:

$$p = \frac{\sum_{c\in C, c\leq\eta^T y} \tilde{\phi}(c;\gamma,\eta^T\Sigma\eta)}{\sum_{c\in C} \tilde{\phi}(c;\gamma,\eta^T\Sigma\eta)}$$

$[-m, m]$. *Let* $U \subset \mathbf{R}$ *be an open interval, and* p_m *be the p-value from Algorithm 2 using* D_m. *We have*

$$\lim_{m\to\infty} p_m = F(\eta^T y; \gamma, U(H, y_0, \frac{\Sigma\eta}{\eta^T\Sigma\eta})).$$

3.4 Inference in the full model

In Chapter 2, we focused on inference for the submodel coefficients $\beta^\star_M = X^+_M\mu$. In selective inference, the choice of the model M is selected via an algorithm *e.g.* the lasso, and the COS method constructed confidence intervals

$$\mathbb{P}\left(\beta^\star_{j,\hat{M}} \in C_j\right) = 1 - \alpha.$$

One possible criticism of the selective confidence intervals for submodel coefficients is the interpretability of the quantity $\beta^\star_{j,\hat{M}}$, since this is the population regression coefficient of variable j within the model $\hat{M}$. The significance of variable j depends on the choice of model meaning variable j can be significant in model M_1, but not significant in M_2, which makes interpretation difficult.

However, this is not an inherent limitation of the COS method. As we saw in the previous two sections, the COS method is not specific to the submodel coefficients. We simply need to change the space of hypothesis $\mathcal{H}$ and the selection function H to perform inference for other regression coefficients.

In many scientific applications, the quantity of interest is the regression coefficient within

the full model $M = [1 \dots p]$. We first discuss the case of $n \geq p$. Let us assume that $y \sim \mathcal{N}(\mu, \sigma^2 I)$. In ordinary least squares , the parameter of interest is $\beta^0 = X^+\mu$, and a classical confidence interval guarantees

$$\mathbb{P}(\beta_j^0 \in C_j) = 1 - \alpha.$$

In the case of least squares after variable selection, we only want to make a confidence interval for the $j \in \hat{M}$, or variables selected by the lasso. This corresponds to inference for a subset $\beta^0_{\hat{M}} = E_{\hat{M}}\beta^0$, where E_M selects the coordinates in M. The interpretation of $\beta^0_{\hat{M}}$ for $j \in \hat{M}$ is clear; this is the regression coefficient of the least squares coefficient restricted to the set selected by the lasso.

For each coefficient $j \in \hat{M}$, Equation (2.5.1) provides a valid p-value of the hypothesis $\beta^0_{j,\hat{M}} = \gamma$,

$$p_j = F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\gamma,\ \sigma^2||\eta_j||_2^2}(\hat{\beta}_{j,\hat{M}}), \tag{3.4.1}$$

where $\eta_j = (X^T_{\hat{M}})^+ e_j$. By inverting, we obtain a selective confidence interval

$$\mathbb{P}\left(\beta^0_{j,\hat{M}} \in C_j\right) = 1 - \alpha. \tag{3.4.2}$$

3.4.1 False Discovery Rate

In this section, we show how to combine selective confidence intervals with the Benjamini-Yeuketieli procedure for FDR control. False discovery rate (FDR) is defined as,

$$\mathbf{E}\left[\frac{V}{R}\right],$$

where V is the number of incorrectly rejected hypotheses and R is the total number of rejected hypotheses. We will restrict ourselves to the case of the well-specified linear model, $y = X\beta^0 + \epsilon$, and $n \geq p$ with X having full rank. In the context of linear regression, there is a sequence of hypotheses $H_{0,j} : \beta_j^0 = 0$ and a hypothesis is considered to be incorrectly rejected if $H_{0,j}$ is true, yet the variable is selected.

Given p-values, we can now apply the Benjamini-Yekutieli procedure (Benjamini et al., 2001) for FDR control. Let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(|\hat{M}|)}$ be the order statistics, and $h_{|\hat{M}|} =$

$\sum_{i=1}^{|\hat{M}|} \frac{1}{i}$. Let k be

$$k = \max\left\{k : p_{(k)} \leq \frac{k}{|\hat{M}|h_{|\hat{M}|}}\alpha\right\}, \tag{3.4.3}$$

then reject $p_{(1)}, \ldots, p_{(k)}$.

Theorem 3.4.1. *Consider the procedure that forms p-values using Equation (3.4.1), chooses k via Equation (3.4.3), and rejects $p_{(1)}, \ldots, p_{(k)}$. Then FDR is controlled at level α.*

Proof. Conditioned on the event that variable j is in the lasso active set, $j \in \hat{M}$, then p_j is uniformly distributed among the null variables. Applying the Benjamini-Yekutieli procedure to the p-values $p_{(1)}, \ldots, p_{(\hat{M})}$ guarantees FDR, since the null p-values are uniformly distributed. □

3.4.2 Intervals for coefficients in full model

In this section, we present a method for selective inference for coordinates of the full-model parameter β^0. We will assume the sparse linear model, namely,

$$y = X\beta^0 + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2$ and β^0 is s-sparse. Since $n < p$, we cannot use the method in the previous section since $\beta^0 \neq X^+X\beta^0$. Instead, we will construct a quantity β^d that is extremely close to β^0 and show that $\beta_j^d = \eta_j^T(X\beta^0) + h_j$. We do this by constructing a population version of the debiased estimator.

The debiased estimator presented in Javanmard and Montanari (2013) is

$$\begin{aligned}
\hat{\beta}^d &= \hat{\beta} + \frac{1}{n}\hat{\Theta}X^T(y - X\hat{\beta}) \\
&= \frac{1}{n}\hat{\Theta}X^T y + (I - \hat{\Theta}\hat{\Sigma})\hat{\beta} \\
&= \frac{1}{n}\hat{\Theta}X^T y + (I - \hat{\Theta}\hat{\Sigma}) \begin{bmatrix} \frac{1}{n}\hat{\Sigma}_{\hat{M}}^{-1} X_{\hat{M}}^T y - \lambda\hat{\Sigma}_{\hat{M}}^{-1} s_{\hat{M}} \\ 0 \end{bmatrix}
\end{aligned}$$

where $\hat{\Sigma}_{\hat{M}} := \frac{1}{n} X_{\hat{M}}^T X_{\hat{M}}$ and $\hat{\Theta}$ is an approximate inverse covariance that is the solution to

$$\min \sum_j \hat{\Theta}_j^T \hat{\Sigma} \hat{\Theta}_j$$
$$\text{subject to } \left\| \hat{\Sigma}\hat{\Theta} - I \right\|_\infty \leq C\sqrt{\frac{\log p}{n}}.$$

Define the population quantity β^d by replacing all occurrences of y with μ:

$$\begin{aligned}
\beta^d(M,s) &:= \frac{1}{n}\hat{\Theta}X^T\mu + (I - \hat{\Theta}\hat{\Sigma}) \begin{bmatrix} \frac{1}{n}\hat{\Sigma}_M^{-1} X_M^T \mu - \lambda \hat{\Sigma}_M^{-1} s \\ 0 \end{bmatrix} \\
&= \beta^0 + \frac{1}{n}\hat{\Theta}X^T\mu + (I - \hat{\Theta}\hat{\Sigma}) \begin{bmatrix} \frac{1}{n}\hat{\Sigma}_M^{-1} X_M^T \mu - \lambda \hat{\Sigma}_M^{-1} s - \beta_M^0 \\ -\beta_{-M}^0 \end{bmatrix} \\
&= \left(\frac{1}{n}\Theta X^T + (I - \Theta\hat{\Sigma}) F_M \hat{\Sigma}_M^{-1} X_M^T \right) \mu - \lambda (I - \Theta\hat{\Sigma}) F_M \hat{\Sigma}_M^{-1} s \\
&:= B\mu + h,
\end{aligned} \tag{3.4.4}$$

where F_M is the matrix such that it takes an $|M|$ vector and pads with 0 to make a p vector.

By choosing η as a row of B, COS framework provides a selective test and confidence interval,

$$H_0 : \beta_j^d(\hat{M}, \hat{s}) = \gamma - \eta^T h$$
$$\mathbb{P}(\beta_j^d(\hat{M}, \hat{s}) \in C_j) = 1 - \alpha.$$

The next step is to show that $\beta^d(\hat{M}, \hat{s})$ is close to β^0, so by appropriately widening C_j, we cover β^0.

Theorem 3.4.2. *Assume that lasso is consistent in the sense* $\left\| \hat{\beta} - \beta^0 \right\|_1 \leq c_L s\sqrt{\frac{\log p}{n}}$, Θ *satisfies* $\left\| \Theta\hat{\Sigma} - I \right\|_\infty \leq c_\Theta \sqrt{\frac{\log p}{n}}$, *and* X *has the sparse eigenvalue condition* $\mu(S,k) := \min_{\|v\|_0 \leq k, \|v\|_2 = 1} \frac{1}{n} \left\| v^T S v \right\| > 0$, *and the empirical sparsity* $\hat{s} := |\hat{M}| < c_M s$, *then*

$$\left\| \beta^d(\hat{M}, \hat{s}) - \beta^0 \right\|_\infty \leq C_{\beta^d} \frac{s \log p}{n}.$$

Proof. Starting from Equation (3.4.4), we have

$$\begin{aligned}
\beta^d - \beta^0 &= (I - \Theta\hat{\Sigma}) \begin{bmatrix} \frac{1}{n}\hat{\Sigma}_{\hat{M}}^{-1} X_{\hat{M}}^T \mu - \lambda \hat{\Sigma}_{\hat{M}}^{-1} s - \beta_{\hat{M}}^0 \\ -\beta_{-\hat{M}}^0 \end{bmatrix} \\
\left\|\beta^d - \beta^0\right\|_\infty &\leq \left\|I - \Theta\hat{\Sigma}\right\|_\infty \left\| \begin{bmatrix} \frac{1}{n}\hat{\Sigma}_{\hat{M}}^{-1} X_{\hat{M}}^T \mu - \lambda \hat{\Sigma}_{\hat{M}}^{-1} s - \beta_{\hat{M}}^0 \\ -\beta_{-\hat{M}}^0 \end{bmatrix} \right\|_1 \\
&\leq c_\Theta \sqrt{\frac{\log p}{n}} \left(\left\| \frac{1}{n}\hat{\Sigma}_{\hat{M}}^{-1} X_{\hat{M}}^T \mu - \lambda \hat{\Sigma}_{\hat{M}}^{-1} s - \beta_{\hat{M}}^0 \right\|_1 + \left\| \beta_{-\hat{M}}^0 \right\|_1 \right) \\
&\leq c_\Theta \sqrt{\frac{\log p}{n}} \left(\left\| \frac{1}{n}\hat{\Sigma}_{\hat{M}}^{-1} X_{\hat{M}}^T \mu - \lambda \hat{\Sigma}_{\hat{M}}^{-1} s - \beta_{\hat{M}}^0 \right\|_1 + \left\| \hat{\beta} - \beta^0 \right\|_1 \right) \\
&\leq c_\Theta \sqrt{\frac{\log p}{n}} \left(\left\| \frac{1}{n}\hat{\Sigma}_{\hat{M}}^{-1} X_{\hat{M}}^T \mu - \lambda \hat{\Sigma}_{\hat{M}}^{-1} s - \beta_{\hat{M}}^0 \right\|_1 + c_L s \sqrt{\frac{\log p}{n}} \right)
\end{aligned}$$

where we used the lasso consistency assumption, $\left\|\Theta\hat{\Sigma} - I\right\|_\infty \leq c_\Theta \sqrt{\frac{\log p}{n}}$n, and the second to last inequality uses the fact that $\hat{\beta}_{-\hat{M}} = 0$, so $\left\|\hat{\beta} - \beta^0\right\|_1 = \left\|\hat{\beta}_{\hat{M}} - \beta_{\hat{M}}^0\right\|_1 + \left\|-\beta_{-\hat{M}}^0\right\|_1 \geq \left\|-\beta_{-\hat{M}}^0\right\|_1$.

We now show $\left\|\frac{1}{n}\hat{\Sigma}^{-1}_{\hat{M}}X^T_{\hat{M}}\mu - \lambda\hat{\Sigma}^{-1}_{\hat{M}}s - \beta^0_{\hat{M}}\right\|_1 \leq s\sqrt{\frac{\log p}{n}}$.

$$\begin{aligned}
&\left\|\frac{1}{n}\hat{\Sigma}^{-1}_{\hat{M}}X^T_{\hat{M}}\mu - \lambda\hat{\Sigma}^{-1}_{\hat{M}}s - \beta^0_{\hat{M}}\right\|_1 \leq \left\|(\frac{1}{n}\hat{\Sigma}^{-1}_{\hat{M}}X^T_{\hat{M}}\mu - \lambda\hat{\Sigma}^{-1}_{\hat{M}}s) - \hat{\beta}_{\hat{M}}\right\|_1 + \left\|\hat{\beta}_{\hat{M}} - \beta^0_{\hat{M}}\right\|_1 \\
&\leq \left\|(\frac{1}{n}\hat{\Sigma}^{-1}_{\hat{M}}X^T_{\hat{M}}\mu - \lambda\hat{\Sigma}^{-1}_{\hat{M}}s) - (\frac{1}{n}\hat{\Sigma}^{-1}_{\hat{M}}X^T_{\hat{M}}y - \lambda\hat{\Sigma}^{-1}_{\hat{M}}s)\right\|_1 + c_L s\sqrt{\frac{\log p}{n}} \\
&\leq \left\|\frac{1}{n}\hat{\Sigma}^{-1}_{\hat{M}}X^T_{\hat{M}}\epsilon\right\|_1 + c_L s\sqrt{\frac{\log p}{n}} \\
&\leq \sqrt{\hat{s}}\left\|\frac{1}{n}\hat{\Sigma}^{-1}_{\hat{M}}X^T_{\hat{M}}\epsilon\right\|_2 + c_L s\sqrt{\frac{\log p}{n}} \\
&\leq \sqrt{\hat{s}}\left\|\hat{\Sigma}^{-1}_{\hat{M}}\right\|_2\left\|\frac{1}{n}X^T_{\hat{M}}\epsilon\right\|_2 + c_L s\sqrt{\frac{\log p}{n}} \\
&\leq \sqrt{\hat{s}}\left\|\hat{\Sigma}^{-1}_{\hat{M}}\right\|_2\sqrt{\hat{s}}\left\|\frac{1}{n}X^T\epsilon\right\|_\infty + c_L s\sqrt{\frac{\log p}{n}} \\
&\leq \hat{s}\sqrt{\frac{\log p}{n}}\left\|\hat{\Sigma}^{-1}_{\hat{M}}\right\|_2 + c_L s\sqrt{\frac{\log p}{n}} \\
&\leq \frac{1}{\lambda_{\min}(\frac{1}{n}X^T_{\hat{M}}X_{\hat{M}})}\hat{s}\sqrt{\frac{\log p}{n}} + c_L s\sqrt{\frac{\log p}{n}} \\
&\leq \left(\frac{1}{\mu(\hat{\Sigma}, c_M s)}c_M + c_L\right)s\sqrt{\frac{\log p}{n}}
\end{aligned}$$

, where $\hat{s} = |\hat{M}|$,and $\lambda_{\min}(\frac{1}{n}X^T_{\hat{M}}X_{\hat{M}}) \geq \mu(\hat{s}) > \mu(c_M s)$.

Plugging this into the expression for $\left\|\beta^d - \beta^0\right\|$,

$$\left\|\beta^d - \beta^0\right\|_\infty \leq c_\Theta\sqrt{\frac{\log p}{n}}\left(\left(\frac{1}{\mu(\hat{\Sigma}, c_M s)}c_M + c_L\right)s\sqrt{\frac{\log p}{n}} + c_L s\sqrt{\frac{\log p}{n}}\right) \tag{3.4.5}$$

$$\leq \left(\frac{c_\Theta}{\mu(\hat{\Sigma}, c_M s)}c_M + 2c_L c_\Theta\right)\frac{s\log p}{n} \tag{3.4.6}$$

□

Lemma 3.4.3 (Assumptions hold under random Gaussian design with additive Gaussian noise)**.** *Assume that the rows of $X \sim \mathcal{N}(0, \Sigma)$ and $n \gg c_s s\log p$. Then the estimation consistency property, existence of a good approximation $\hat{\Theta}$, empirical sparsity $\hat{s} < c_M s$, and $\mu(\hat{\Sigma}, c_M s) > \frac{1}{2}\mu(\Sigma, c_M s)$ with probability tending to 1.*

Proof. The estimation consistency property follows from Negahban et al. (2012). The bound $\left\|\hat{\Sigma}\hat{\Theta} - I\right\|_\infty < \sqrt{\frac{\log p}{n}}$ is established in Javanmard and Montanari (2013). The empirical sparsity result $\hat{s} \leq c_M s$ is from Belloni et al. (2011, 2013).

The condition on concentration of sparse eigenvalues can be derived using Loh and Wainwright (2012, Lemma 15, Supplementary Materials). Lemma 15 states if X is a zero-mean sub-Gaussian matrix with covariance Σ and subgaussian parameter σ^2, then there is a universal constant $c > 0$ such that

$$\mathbb{P}\left(\sup_{\|v\|_0 \leq s, \|v\|_2 = 1} |\frac{1}{n} v^T X^T X v - v^T \Sigma v| \geq t\right) \leq 2\exp\left(-cn\min(\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}) + s\log p\right). \quad (3.4.7)$$

With high probability and for all $v \in K(s) := \{v : \|v\|_0 \leq s, \|v\|_2 = 1\}$,

$$\begin{aligned}
|v^T\hat{\Sigma}v - v^T\Sigma v| &< t \\
v^T\Sigma v - t &< v^T\hat{\Sigma}v \\
\min_{w \in K(s)} w^T\Sigma w - t &< v^T\hat{\Sigma}v \\
\mu(\Sigma, s) - t &< v^T\hat{\Sigma}v \\
\mu(\Sigma, s) - t &< \min_{v \in K(s)} v^T\hat{\Sigma}v \\
\mu(\Sigma, s) - t &< \mu(\hat{\Sigma}, s).
\end{aligned}$$

We now use this to show $\mu(\hat{\Sigma}, C_m s) > \frac{1}{2}\mu(\Sigma, c_M s)$. Let $t = \frac{1}{2}\mu(\Sigma, c_M s)$, then by the previous argument and Equation (3.4.7),

$$\mu(\hat{\Sigma}, c_M s) > \frac{1}{2}\mu(\Sigma, c_M s)$$

with probability at least

$$1 - 2\exp\left(-cn\min(\frac{\mu(\Sigma, c_M s)^2}{4\sigma^4}, \frac{\mu(\Sigma, c_M s)}{2\sigma^2}) + c_M s\log p\right).$$

For $n > \frac{2c_M s\log p}{c\min(\frac{\mu(\Sigma, c_M s)^2}{4\sigma^4}, \frac{\mu(\Sigma, c_M s)}{2\sigma^2})}$, we have with probability at least,

$$1 - 2\exp\left(-\frac{1}{2}cn\min(\frac{\mu(\Sigma, c_M s)^2}{4\sigma^4}, \frac{\mu(\Sigma, c_M s)}{2\sigma^2})\right).$$

□

Corollary 3.4.4. *Under the assumptions of Lemma 3.4.3, and* $n \gg s^2 \log^2 p$,

$$\left\|\beta^d(\hat{M}, \hat{s}) - \beta^0\right\|_\infty \leq \frac{\delta}{\sqrt{n}}$$

for any $\delta > 0$.

Proof. Lemma 3.4.3 ensures that $\mu(\hat{\Sigma}, s) > \frac{1}{2}\mu(\Sigma, s)$, and plugging this into Equation (3.4.6) gives

$$\left\|\beta^d - \beta^0\right\|_\infty \leq \left(\frac{2c_\Theta}{\mu(\Sigma, c_M s)} c_M + 2c_L c_\Theta\right) \frac{s \log p}{n}$$

Since $n \gg s^2 \log^2 p$, we have

$$\begin{aligned}\left\|\beta^d - \beta^0\right\|_\infty &\leq \left(\frac{2c_\Theta}{\mu(\Sigma, c_M s)} c_M + 2c_L c_\Theta\right) \frac{s \log p}{n} \\ &\leq \left(\frac{2c_\Theta}{\mu(\Sigma, c_M s)} c_M + 2c_L c_\Theta\right) o(\frac{1}{\sqrt{n}}) \\ &\leq \frac{\delta}{\sqrt{n}}.\end{aligned}$$

□

Corollary 3.4.5. *Let* C_j *be a selective confidence interval for* β^d *meaning* $\mathbb{P}(\beta_j^d \in C_j) = 1 - \alpha$, *then*

$$\liminf \mathbb{P}(\beta_j^0 \in C_j \pm \frac{\delta}{\sqrt{n}}) \geq 1 - \alpha.$$

Proof. With probability at least $1 - \alpha$, $\beta_j^d \in C_j$ and with probability tending to $1 - o(1)$, $\beta_j^d - \beta_j^0 < \frac{\delta}{\sqrt{n}}$. Thus with probability at least $1 - \alpha - o(1)$, $\beta_j^0 \in C_j \pm \frac{\delta}{\sqrt{n}}$. □

Figure 3.1 shows the results of a simulation study. It makes clear that the intervals of Javanmard and Montanari (2013) and our selective confidence intervals cover β^d, which is close to β^0. The Javanmard-Montanari intervals are the high-dimensional analog of a z-interval, so they are not selectively valid, unlike the selective intervals in blue.

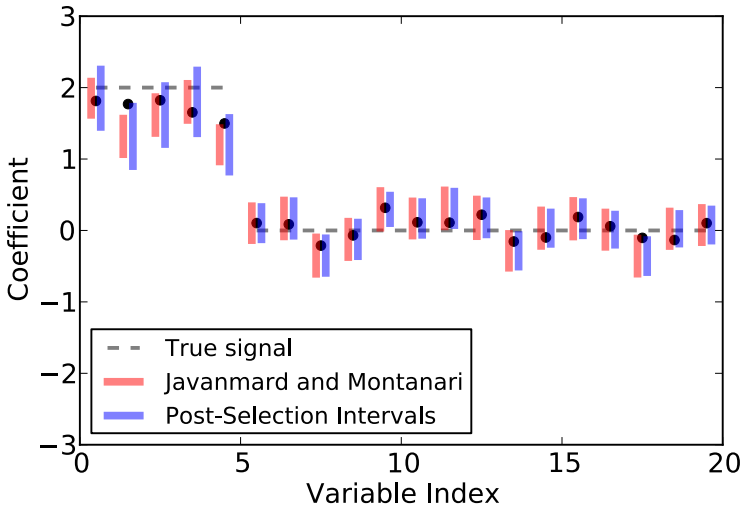


Figure 3.1: Confidence intervals for the coefficients in a design with $n = 25$, $p = 50$, and 5 non-zero coefficients. Only the first 20 coefficients are shown. The dotted line represents the true signal, and the points represent the (biased) post-selection target. The colored bars denote the intervals.

3.5 Selective Inference for the Knockoff Filter

In this section, we show how to make selectively valid confidence intervals for the knockoff method. Let $\tilde{X}$ be the knockoff design matrix, so the knockoff regression is done on $y = [X; \tilde{X}]\beta + \epsilon$. The introduction of the knockoff variables, $\tilde{X}$, allows us to estimate the FDP as the number of knockoff variables selected divided by the number of true variables selected:

$$FDP(M) = \frac{|M \cap \tilde{X}|}{|M \cap X| \vee 1} \tag{3.5.1}$$

Given a sequence of models $M(1), \ldots, M(k)$ be a sequence of nested models $M(k) \subset M_(k-1) \subset \ldots \subset M(1) \subset [1 \,..\, 2p]$. For the lasso, where the $M(j)$ correspond to the lasso active set at λ_j, the models are not necessarily nested. We define $M(j) = \cup_{l=k}^{j} A_j$, where A_j is the active set of lasso at λ_j. We have an estimate FDP estimate for each model, $FDP(M(j)) = \frac{|M(j) \cap \tilde{X}|}{|M(j) \cap X|}$, where X and $\tilde{X}$ represent the indices of the real and knockoff variables respectively. This suggests selecting the largest model such that the FDP estimate

is less than α,

$$T = \min\{t : FDP(M(t)) \leq \alpha\}. \tag{3.5.2}$$

To show this controls modified FDR, we need to construct W-statistics such that our stopping rule, corresponds to the stopping rule of Foygel Barber and Candes (2014).

Theorem 3.5.1. *The model selected by the stopping rule in Equation (3.5.2) controls the modified FDR, that is*

$$\mathbb{E}\left[\frac{|M(T) \cap V|}{|M(T) \cap X| + 1/\alpha}\right] \leq \alpha$$

where $V = \{1 \leq j \leq p : \beta_j = 0\}$.

Theorem 3.5.2. *We construct some W statistics. Define* $t_j = \min\{t : x_j \in M(t)\}$ *and* $\tilde{t}_j = \min\{t : \tilde{x}_j \in M(t)\}$. *Define*

$$W_j = \begin{cases} t_j & \text{if } t_j < \tilde{t}_j \\ -\tilde{t}_j & \text{if } \tilde{t}_j < t_j \end{cases}$$

We now verify that the FDP estimate given in Foygel Barber and Candes (2014) using the W-statistics are the same FDP estimate as (3.5.1).

$$\begin{aligned} FDP_W(t) &= \frac{|\{j : W_j \leq -t\}|}{|\{j : W_j > t\} \vee 1} \\ &= \frac{\{j : \tilde{t}_j > t, j \in \tilde{X}\}}{|\{j : t_j > t, j \in X\}| \vee 1} \\ &= \frac{\{j : M(t) \cap \tilde{X}\}}{|\{j : M(t) \cap X\}| \vee 1}. \end{aligned}$$

By invoking the main theorem of Foygel Barber and Candes (2014), we see that (3.5.2) controls the modified FDR.

By using the FDP^+ estimate in place of equation (3.5.2),

$$FDP^+(M) = \frac{|M \cap \tilde{X}|}{|M \cap X| \vee 1 + 1} \tag{3.5.3}$$

$$T^+ = \min\{t : FDP^+(M(t)) \leq \alpha\}. \tag{3.5.4}$$

we can control FDR, instead of modified FDR.

Theorem 3.5.3. *The model selected by the stopping rule in Equation (3.5.4) controls FDR, that is*

$$\mathbb{E}\left[\frac{|M(T) \cap V|}{|M(T) \cap X| \vee 1}\right] \leq \alpha$$

where $V = \{1 \leq j \leq p : \beta_j = 0\}$.

Proof. Same as the previous theorem. □

Let $M^\star = KO(y)$ be the final model returned by the knockoff procedure k applied to the regression pair (y, X) using the lasso models at the sequence $\lambda_1, \ldots, \lambda_k$. Our goal is to do inference for $\beta_j^0 = e_j X^+ \mu$ for some $j \in M^\star$. The selection event, the set of y's that lead us to testing β_j^0, is $S_j = \{y : j \in KO(y)\}$. This precise set is difficult to analytically describe, so we resort to Algorithm 2.

We can analytically describe the finer event

$$S = \{y : (L(y, \lambda_1), \ldots, L(y, \lambda_{T+1})) = (M(1), \ldots, M(T+1))\} \subset S_j,$$

where $L(y, \lambda)$ is the active set of lasso at λ. For any $y \in S$, the knockoff procedure defined by the stopping rule (3.5.2) returns the same set of variables, so $S \subset S_j$. The set S is described by the intersection of the union of linear inequalities given in Section 2.7. This allows us to do inference using the results of Theorem 2.4.3.

We next describe a method using the general method of Chapter 3.3. Using the COS method, we first describe the knockoff selection event. The selection event for variable j is $S_j = \{y : j \in KO(y)\}$. The general method instead uses the one-dimensional finer selection event $U_j = \{c : j \in KO(P^\perp_{\Sigma,\eta} y + c\frac{\Sigma\eta}{\eta^T \Sigma \eta})\}$. This set is approximated using Algorithm 2 that computes an approximation to U_j and an approximate p-value.

Since the knockoff method assumes a well-specified linear model, we can use the reference distribution $y \sim \mathcal{N}(X\beta, \sigma^2 I)$ instead of $y \sim \mathcal{N}(\mu, \sigma^2 I)$. This is the well-specified linear regression model of Fithian et al. (2014). The selection event is now $U_j = \{C : j \in KO(P^\perp_{X_{-j}} y + C), C \in span(X_{-j})^\perp\}$. A multi-dimensional analog of Algorithm 2 can now be applied, but the search set D is now over a $n - p + 1$ dimensional subset.

Part II

Learning Mixed Graphical Models

Chapter 4

Learning Mixed Graphical Models

4.1 Introduction

Many authors have considered the problem of learning the edge structure and parameters of sparse undirected graphical models. We will focus on using the l_1 regularizer to promote sparsity. This line of work has taken two separate paths: one for learning continuous valued data and one for learning discrete valued data. However, typical data sources contain both continuous and discrete variables: population survey data, genomics data, url-click pairs etc. For genomics data, in addition to the gene expression values, we have attributes attached to each sample such as gender, age, ethniticy etc. In this work, we consider learning mixed models with both continuous Gaussian variables and discrete categorical variables.

For only continuous variables, previous work assumes a multivariate Gaussian (Gaussian graphical) model with mean 0 and inverse covariance Θ. Θ is then estimated via the graphical lasso by minimizing the regularized negative log-likelihood $\ell(\Theta) + \lambda \|\Theta\|_1$. Several efficient methods for solving this can be found in Friedman et al. (2008a); Banerjee et al. (2008). Because the graphical lasso problem is computationally challenging, several authors considered methods related to the pseudolikelihood (PL) and nodewise regression (Meinshausen and Bühlmann, 2006; Friedman et al., 2010a; Peng et al., 2009). For discrete models, previous work focuses on estimating a pairwise Markov random field of the form $p(y) \propto \exp \sum_{r \leq j} \phi_{rj}(y_r, y_j)$, where ϕ_{rj} are pairwise potentials. The maximum likelihood problem is intractable for models with a moderate to large number of variables (high-dimensional) because it requires evaluating the partition function and its derivatives. Again previous work has focused on the pseudolikelihood approach (Guo et al., 2010; Schmidt,

2010; Schmidt et al., 2008; Höfling and Tibshirani, 2009; Jalali et al., 2011; Lee et al., 2006; Ravikumar et al., 2010).

Our main contribution here is to propose a model that connects the discrete and continuous models previously discussed. The conditional distributions of this model are two widely adopted and well understood models: multiclass logistic regression and Gaussian linear regression. In addition, in the case of only discrete variables, our model is a pairwise Markov random field; in the case of only continuous variables, it is a Gaussian graphical model. Our proposed model leads to a natural scheme for structure learning that generalizes the graphical Lasso. Here the parameters occur as singletons, vectors or blocks, which we penalize using group-lasso norms, in a way that respects the symmetry in the model. Since each parameter block is of different size, we also derive a calibrated weighting scheme to penalize each edge fairly. We also discuss a conditional model (conditional random field) that allows the output variables to be mixed, which can be viewed as a multivariate response regression with mixed output variables. Similar ideas have been used to learn the covariance structure in multivariate response regression with continuous output variables Witten and Tibshirani (2009); Kim et al. (2009); Rothman et al. (2010).

In Section 4.2, we introduce our new mixed graphical model and discuss previous approaches to modeling mixed data. Section 4.3 discusses the pseudolikelihood approach to parameter estimation and connections to generalized linear models. Section 4.4 discusses a natural method to perform structure learning in the mixed model. Section 4.5 presents the calibrated regularization scheme, Section 4.6 discusses the consistency of the estimation procedures, and Section 4.7 discusses two methods for solving the optimization problem. Finally, Section 4.8 discusses a conditional random field extension and Section 4.9 presents empirical results on a census population survey dataset and synthetic experiments.

4.2 Mixed Graphical Model

We propose a pairwise graphical model on continuous and discrete variables. The model is a pairwise Markov random field with density $p(x, y; \Theta)$ proportional to

$$\exp\left(\sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \sum_{s=1}^{p}\alpha_s x_s + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j)x_s + \sum_{j=1}^{q}\sum_{r=1}^{q}\phi_{rj}(y_r, y_j)\right). \quad (4.2.1)$$

Here x_s denotes the sth of p continuous variables, and y_j the jth of q discrete variables. The joint model is parametrized by $\Theta = [\{\beta_{st}\}, \{\alpha_s\}, \{\rho_{sj}\}, \{\phi_{rj}\}]$. The discrete y_r takes on L_r states. The model parameters are β_{st} continuous-continuous edge potential, α_s continuous node potential, $\rho_{sj}(y_j)$ continuous-discrete edge potential, and $\phi_{rj}(y_r, y_j)$ discrete-discrete edge potential. $\rho_{sj}(y_j)$ is a function taking L_j values $\rho_{sj}(1), \ldots, \rho_{sj}(L_j)$. Similarly, $\phi_{rj}(y_r, y_j)$ is a bivariate function taking on $L_r \times L_j$ values. Later, we will think of $\rho_{sj}(y_j)$ as a vector of length L_j and $\phi_{rj}(y_r, y_j)$ as a matrix of size $L_r \times L_j$.

The two most important features of this model are:

1. the conditional distributions are given by Gaussian linear regression and multiclass logistic regressions;

2. the model simplifies to a multivariate Gaussian in the case of only continuous variables and simplifies to the usual discrete pairwise Markov random field in the case of only discrete variables.

The conditional distributions of a graphical model are of critical importance. The absence of an edge corresponds to two variables being conditionally independent. The conditional independence can be read off from the conditional distribution of a variable on all others. For example in the multivariate Gaussian model, x_s is conditionally independent of x_t iff the partial correlation coefficient is 0. The partial correlation coefficient is also the regression coefficient of x_t in the linear regression of x_s on all other variables. Thus the conditional independence structure is captured by the conditional distributions via the regression coefficient of a variable on all others. Our mixed model has the desirable property that the two type of conditional distributions are simple Gaussian linear regressions and multiclass logistic regressions. This follows from the pairwise property in the joint distribution. In more detail:

1. The conditional distribution of y_r given the rest is multinomial, with probabilities defined by a multiclass logistic regression where the covariates are the other variables x_s and $y_{\setminus r}$ (denoted collectively by z in the right-hand side):

$$p(y_r = k | y_{\setminus r}, x; \Theta) = \frac{\exp\left(\omega_k^T z\right)}{\sum_{l=1}^{L_r} \exp\left(\omega_l^T z\right)} = \frac{\exp\left(\omega_{0k} + \sum_j \omega_{kj} z_j\right)}{\sum_{l=1}^{L_r} \exp\left(\omega_{0l} + \sum_j \omega_{lj} z_j\right)} \tag{4.2.2}$$

   Here we use a simplified notation, which we make explicit in Section 4.3.1. The discrete

variables are represented as dummy variables for each state, e.g. $z_j = \mathbb{1}[y_u = k]$, and for continuous variables $z_s = x_s$.

2. The conditional distribution of x_s given the rest is Gaussian, with a mean function defined by a linear regression with predictors $x_{\setminus s}$ and y_r.

$$E(x_s|x_{\setminus s}, y_r; \Theta) = \omega^T z = \omega_0 + \sum_j z_j \omega_j \tag{4.2.3}$$
$$p(x_s|x_{\setminus s}, y_r; \Theta) = \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{1}{2\sigma_s^2}(x_s - \omega^T z)^2\right).$$

As before, the discrete variables are represented as dummy variables for each state $z_j = \mathbb{1}[y_u = k]$ and for continuous variables $z_s = x_s$.

The exact form of the conditional distributions (4.2.2) and (4.2.3) are given in (4.3.5) and (4.3.4) in Section 4.3.1, where the regression parameters ω_j are defined in terms of the parameters Θ.

The second important aspect of the mixed model is the two special cases of only continuous and only discrete variables.

1. Continuous variables only. The pairwise mixed model reduces to the familiar multivariate Gaussian parametrized by the symmetric positive-definite inverse covariance matrix $B = \{\beta_{st}\}$ and mean $\mu = B^{-1}\alpha$,

$$p(x) \propto \exp\left(-\frac{1}{2}(x - B^{-1}\alpha)^T B(x - B^{-1}\alpha)\right).$$

2. Discrete variables only. The pairwise mixed model reduces to a pairwise discrete (second-order interaction) Markov random field,

$$p(y) \propto \exp\left(\sum_{j=1}^{q}\sum_{r=1}^{q} \phi_{rj}(y_r, y_j)\right).$$

Although these are the most important aspects, we can characterize the joint distribution further. The conditional distribution of the continuous variables given the discrete follow a multivariate Gaussian distribution, $p(x|y) = \mathcal{N}(\mu(y), B^{-1})$. Each of these Gaussian distributions share the same inverse covariance matrix B but differ in the mean parameter,

since all the parameters are pairwise. By standard multivariate Gaussian calculations,

$$p(x|y) = \mathcal{N}(B^{-1}\gamma(y), B^{-1}) \tag{4.2.4}$$

$$\{\gamma(y)\}_s = \alpha_s + \sum_j \rho_{sj}(y_j) \tag{4.2.5}$$

$$p(y) \propto \exp\left(\sum_{j=1}^{q}\sum_{r=1}^{j}\phi_{rj}(y_r, y_j) + \frac{1}{2}\gamma(y)^T B^{-1}\gamma(y)\right) \tag{4.2.6}$$

Thus we see that the continuous variables conditioned on the discrete are multivariate Gaussian with common covariance, but with means that depend on the value of the discrete variables. The means depend additively on the values of the discrete variables since $\{\gamma(y)\}_s = \sum_{j=1}^{r}\rho_{sj}(y_j)$. The marginal $p(y)$ has a known form, so for models with few number of discrete variables we can sample efficiently.

4.2.1 Related work on mixed graphical models

Lauritzen (1996) proposed a type of mixed graphical model, with the property that conditioned on discrete variables, $p(x|y) = \mathcal{N}(\mu(y), \Sigma(y))$. The homogeneous mixed graphical model enforces common covariance, $\Sigma(y) \equiv \Sigma$. Thus our proposed model is a special case of Lauritzen's mixed model with the following assumptions: common covariance, additive mean assumptions and the marginal $p(y)$ factorizes as a pairwise discrete Markov random field. With these three assumptions, the full model simplifies to the mixed pairwise model presented. Although the full model is more general, the number of parameters scales exponentially with the number of discrete variables, and the conditional distributions are not as convenient. For each state of the discrete variables there is a mean and covariance. Consider an example with q binary variables and p continuous variables; the full model requires estimates of 2^q mean vectors and covariance matrices in p dimensions. Even if the homogeneous constraint is imposed on Lauritzen's model, there are still 2^q mean vectors for the case of binary discrete variables. The full mixed model is very complex and cannot be easily estimated from data without some additional assumptions. In comparison, the mixed pairwise model has number of parameters $O((p+q)^2)$ and allows for a natural regularization scheme which makes it appropriate for high dimensional data.

An alternative to the regularization approach that we take in this paper, is the limited-order correlation hypothesis testing method Tur and Castelo (2012). The authors develop a

hypothesis test via likelihood ratios for conditional independence. However, they restrict to the case where the discrete variables are marginally independent so the maximum likelihood estimates are well-defined for $p > n$.

There is a line of work regarding parameter estimation in undirected mixed models that are decomposable: any path between two discrete variables cannot contain only continuous variables. These models allow for fast exact maximum likelihood estimation through node-wise regressions, but are only applicable when the structure is known and $n > p$ (Edwards, 2000). There is also related work on parameter learning in directed mixed graphical models. Since our primary goal is to learn the graph structure, we forgo exact parameter estimation and use the pseudolikelihood. Similar to the exact maximum likelihood in decomposable models, the pseudolikelihood can be interpreted as node-wise regressions that enforce symmetry.

After we proposed our model[1], the independent work of Cheng et al. (2013) appeared, which considers a more complicated mixed graphical model. Their model includes higher order interaction terms by allowing the covariance of the continuous variables to be a function of the categorical variables y, which results in a larger model similar to Lauritzen's model. This results in a model with $O(p^2q + q^2)$ parameters, as opposed to $O(p^2 + q^2)$ in our proposed pairwise model. We believe that in the high-dimensional setting where data sparsity is an issue a simpler model is an advantage.

To our knowledge, this work is the first to consider convex optimization procedures for learning the edge structure in mixed graphical models.

[1] Cheng et al. (2013), http://arxiv.org/abs/1304.2810, appeared on arXiv 11 months after our original paper was put on arXiv, http://arxiv.org/abs/1205.5012.

4.3 Parameter Estimation: Maximum Likelihood and Pseudolikelihood

Given samples $(x_i, y_i)_{i=1}^n$, we want to find the maximum likelihood estimate of Θ. This can be done by minimizing the negative log-likelihood of the samples:

$$\ell(\Theta) = -\sum_{i=1}^{n} \log p(x_i, y_i; \Theta) \text{ where} \tag{4.3.1}$$

$$\begin{aligned}\log p(x, y; \Theta) = &\sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \sum_{s=1}^{p}\alpha_s x_s + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j)x_s \\ &+ \sum_{j=1}^{q}\sum_{r=1}^{j}\phi_{rj}(y_r, y_j) - \log Z(\Theta)\end{aligned} \tag{4.3.2}$$

The negative log-likelihood is convex, so standard gradient-descent algorithms can be used for computing the maximum likelihood estimates. The major obstacle here is $Z(\Theta)$, which involves a high-dimensional integral. Since the pairwise mixed model includes both the discrete and continuous models as special cases, maximum likelihood estimation is at least as difficult as the two special cases, the first of which is a well-known computationally intractable problem. We defer the discussion of maximum likelihood estimation to the supplementary material.

4.3.1 Pseudolikelihood

The pseudolikelihood method Besag (1975) is a computationally efficient and consistent estimator formed by products of all the conditional distributions:

$$\tilde{\ell}(\Theta|x, y) = -\sum_{s=1}^{p} \log p(x_s|x_{\setminus s}, y; \Theta) - \sum_{r=1}^{q} \log p(y_r|x, y_{\setminus r}; \Theta) \tag{4.3.3}$$

The conditional distributions $p(x_s|x_{\setminus s}, y; \theta)$ and $p(y_r = k|y_{\setminus r,}, x; \theta)$ take on the familiar form of linear Gaussian and (multiclass) logistic regression, as we pointed out in (4.2.2) and (4.2.3). Here are the details:

- The conditional distribution of a continuous variable x_s is Gaussian with a linear

regression model for the mean, and unknown variance.

$$p(x_s|x_{\setminus s}, y; \Theta) = \frac{\sqrt{\beta_{ss}}}{\sqrt{2\pi}} \exp\left(\frac{-\beta_{ss}}{2}\left(\frac{\alpha_s + \sum_j \rho_{sj}(y_j) - \sum_{t\neq s}\beta_{st}x_t}{\beta_{ss}} - x_s\right)^2\right) \quad (4.3.4)$$

- The conditional distribution of a discrete variable y_r with L_r states is a multinomial distribution, as used in (multiclass) logistic regression. Whenever a discrete variable is a predictor, each of its levels contribute an additive effect; continuous variables contribute linear effects.

$$p(y_r|y_{\setminus r,}, x; \Theta) = \frac{\exp\left(\sum_s \rho_{sr}(y_r)x_s + \phi_{rr}(y_r, y_r) + \sum_{j\neq r}\phi_{rj}(y_r, y_j)\right)}{\sum_{l=1}^{L_r}\exp\left(\sum_s \rho_{sr}(l)x_s + \phi_{rr}(l, l) + \sum_{j\neq r}\phi_{rj}(l, y_j)\right)} \quad (4.3.5)$$

Taking the negative log of both gives us

$$-\log p(x_s|x_{\setminus s}, y; \Theta) = -\frac{1}{2}\log\beta_{ss} + \frac{\beta_{ss}}{2}\left(\frac{\alpha_s}{\beta_{ss}} + \sum_j \frac{\rho_{sj}(y_j)}{\beta_{ss}} - \sum_{t\neq s}\frac{\beta_{st}}{\beta_{ss}}x_t - x_s\right)^2 \quad (4.3.6)$$

$$-\log p(y_r|y_{\setminus r,}, x; \Theta) = -\log\frac{\exp\left(\sum_s \rho_{sr}(y_r)x_s + \phi_{rr}(y_r, y_r) + \sum_{j\neq r}\phi_{rj}(y_r, y_j)\right)}{\sum_{l=1}^{L_r}\exp\left(\sum_s \rho_{sr}(l)x_s + \phi_{rr}(l, l) + \sum_{j\neq r}\phi_{rj}(l, y_j)\right)} \quad (4.3.7)$$

A generic parameter block, θ_{uv}, corresponding to an edge (u, v) appears twice in the pseudolikelihood, once for each of the conditional distributions $p(z_u|z_v)$ and $p(z_v|z_u)$.

Proposition 4.3.1. *The negative log pseudolikelihood in (4.3.3) is jointly convex in all the parameters $\{\beta_{ss}, \beta_{st}, \alpha_s, \phi_{rj}, \rho_{sj}\}$ over the region $\beta_{ss} > 0$.*

We prove Proposition 4.3.1 in the Supplementary Materials.

4.3.2 Separate node-wise regression

A simple approach to parameter estimation is via separate node-wise regressions; a generalized linear model is used to estimate $p(z_s|z_{\setminus s})$ for each s. Separate regressions were used in Meinshausen and Bühlmann (2006) for the Gaussian graphical model and Ravikumar et al. (2010) for the Ising model. The method can be thought of as an asymmetric form of the pseudolikelihood since the pseudolikelihood enforces that the parameters are shared across the conditionals. Thus the number of parameters estimated in the separate regression is

approximately double that of the pseudolikelihood, so we expect that the pseudolikelihood outperforms at low sample sizes and low regularization regimes. The node-wise regression was used as our baseline method since it is straightforward to extend it to the mixed model. As we predicted, the pseudolikelihood or joint procedure outperforms separate regressions; see top left box of Figures 4.6 and 4.7. Liu and Ihler (2012, 2011) confirm that the separate regressions are outperformed by pseudolikelihood in numerous synthetic settings.

Concurrent work of Yang et al. (2012, 2013) extend the separate node-wise regression model from the special cases of Gaussian and categorical regressions to generalized linear models, where the univariate conditional distribution of each node $p(x_s|x_{\setminus s})$ is specified by a generalized linear model (e.g. Poisson, categorical, Gaussian). By specifying the conditional distributions, Besag (1974) show that the joint distribution is also specified. Thus another way to justify our mixed model is to define the conditionals of a continuous variable as Gaussian linear regression and the conditionals of a categorical variable as multiple logistic regression and use the results in Besag (1974) to arrive at the joint distribution in (4.2.1). However, the neighborhood selection algorithm in Yang et al. (2012, 2013) is restricted to models of the form $p(x) \propto \exp\left(\sum_s \theta_s x_s + \sum_{s,t} \theta_{st} x_s x_t + \sum_s C(x_s)\right)$. In particular, this procedure cannot be applied to edge selection in our pairwise mixed model in (4.2.1) or the categorical model in (2) with greater than 2 states. Our baseline method of separate regressions is closely related to the neighborhood selection algorithm they proposed; the baseline can be considered as a generalization of Yang et al. (2012, 2013) to allow for more general pairwise interactions with the appropriate regularization to select edges. Unfortunately, the theoretical results in Yang et al. (2012, 2013) do not apply to the baseline nodewise regression method, nor the joint pseudolikelihood.

4.4 Conditional Independence and Penalty Terms

In this section, we show how to incorporate edge selection into the maximum likelihood or pseudolikelihood procedures. In the graphical representation of probability distributions, the absence of an edge $e = (u, v)$ corresponds to a conditional independency statement that variables x_u and x_v are conditionally independent given all other variables (Koller and Friedman, 2009). We would like to maximize the likelihood subject to a penalization on the number of edges since this results in a sparse graphical model. In the pairwise mixed model, there are 3 type of edges

1. β_{st} is a scalar that corresponds to an edge from x_s to x_t. $\beta_{st} = 0$ implies x_s and x_t are conditionally independent given all other variables. This parameter is in two conditional distributions, corresponding to either x_s or x_t is the response variable, $p(x_s|x_{\setminus s}, y; \Theta)$ and $p(x_t|x_{\setminus t}, y; \Theta)$.

2. ρ_{sj} is a vector of length L_j. If $\rho_{sj}(y_j) = 0$ for all values of y_j, then y_j and x_s are conditionally independent given all other variables. This parameter is in two conditional distributions, corresponding to either x_s or y_j being the response variable: $p(x_s|x_{\setminus s}, y; \Theta)$ and $p(y_j|x, y_{\setminus j}; \Theta)$.

3. ϕ_{rj} is a matrix of size $L_r \times L_j$. If $\phi_{rj}(y_r, y_j) = 0$ for all values of y_r and y_j, then y_r and y_j are conditionally independent given all other variables. This parameter is in two conditional distributions, corresponding to either y_r or y_j being the response variable, $p(y_r|x, y_{\setminus r}; \Theta)$ and $p(y_j|x, y_{\setminus j}; \Theta)$.

For conditional independencies that involve discrete variables, the absence of that edge requires that the entire matrix ϕ_{rj} or vector ρ_{sj} is 0 [2]. The form of the pairwise mixed model motivates the following regularized optimization problem

$$\underset{\Theta}{\text{minimize}} \ \ell_\lambda(\Theta) = \ell(\Theta) + \lambda \left(\sum_{s<t} \mathbb{1}[\beta_{st} \neq 0] + \sum_{sj} \mathbb{1}[\rho_{sj} \not\equiv 0] + \sum_{r<j} \mathbb{1}[\phi_{rj} \not\equiv 0] \right). \quad (4.4.1)$$

All parameters that correspond to the same edge are grouped in the same indicator function. This problem is non-convex, so we replace the l_0 sparsity and group sparsity penalties with the appropriate convex relaxations. For scalars, we use the absolute value (l_1 norm), for vectors we use the l_2 norm, and for matrices we use the Frobenius norm. This choice corresponds to the standard relaxation from group l_0 to group l_1/l_2 (group lasso) norm (Bach et al., 2011; Yuan and Lin, 2006),

$$\underset{\Theta}{\text{minimize}} \ \ell_\lambda(\Theta) = \ell(\Theta) + \lambda \left(\sum_{s=1}^{p} \sum_{t=1}^{s-1} |\beta_{st}| + \sum_{s=1}^{p} \sum_{j=1}^{q} \|\rho_{sj}\|_2 + \sum_{j=1}^{q} \sum_{r=1}^{j-1} \|\phi_{rj}\|_F \right). \quad (4.4.2)$$

[2] If $\rho_{sj}(y_j) = \textit{constant}$, then x_s and y_j are also conditionally independent. However, the unpenalized term α will absorb the constant, so the estimated $\rho_{sj}(y_j)$ will never be constant for $\lambda > 0$.

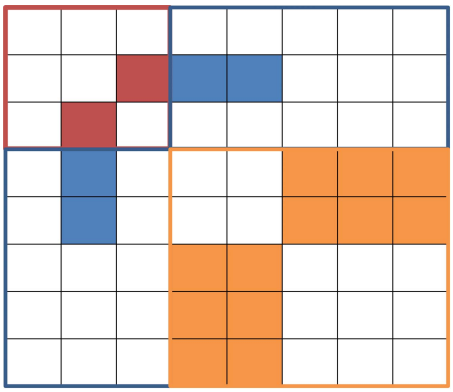


Figure 4.1: *Symmetric matrix represents the parameters Θ of the model. This example has $p = 3$, $q = 2$, $L_1 = 2$ and $L_2 = 3$. The red square corresponds to the continuous graphical model coefficients B and the solid red square is the scalar β_{st}. The blue square corresponds to the coefficients ρ_{sj} and the solid blue square is a vector of parameters $\rho_{sj}(\cdot)$. The orange square corresponds to the coefficients ϕ_{rj} and the solid orange square is a matrix of parameters $\phi_{rj}(\cdot, \cdot)$. The matrix is symmetric, so each parameter block appears in two of the conditional probability regressions.*

4.5 Calibrated regularizers

In (4.4.2) each of the group penalties are treated as equals, irrespective of the size of the group. We suggest a calibration or weighting scheme to balance the load in a more equitable way. We introduce weights for each group of parameters and show how to choose the weights such that each parameter set is treated equally under p_F, the fully-factorized independence model [3]

$$\underset{\Theta}{\text{minimize}}\ \ell(\Theta) + \lambda \left(\sum_{t=1}^{p} \sum_{t=1}^{s-1} w_{st} |\beta_{st}| + \sum_{s=1}^{p} \sum_{j=1}^{q} w_{sj} \left\| \rho_{sj} \right\|_2 + \sum_{j=1}^{q} \sum_{r=1}^{j-1} w_{rj} \left\| \phi_{rj} \right\|_F \right) \tag{4.5.1}$$

Based on the KKT conditions (Friedman et al., 2007), the parameter group θ_g is non-zero if

$$\left\| \frac{\partial \ell}{\partial \theta_g} \right\| > \lambda w_g$$

where θ_g and w_g represents one of the parameter groups and its corresponding weight. Now $\frac{\partial \ell}{\partial \theta_g}$ can be viewed as a generalized residual, and for different groups these are different dimensions—e.g. scalar/vector/matrix. So even under the independence model (when all

[3] Under the independence model p_F is fully-factorized $p(x, y) = \prod_{s=1}^{p} p(x_s) \prod_{r=1}^{q} p(y_r)$

	$\left\|\frac{\partial \ell}{\partial \phi_{12}}\right\|_F$	$\left\|\frac{\partial \ell}{\partial \rho_{11}}\right\|_2$	$\left\|\frac{\partial \ell}{\partial \rho_{21}}\right\|_2$	$\left\|\frac{\partial \ell}{\partial \rho_{12}}\right\|_2$	$\left\|\frac{\partial \ell}{\partial \rho_{22}}\right\|_2$	$\left\|\frac{\partial \ell}{\partial \beta_{12}}\right\|$
Exact weights w_g (4.5.2)	0.18	0.63	0.19	0.47	0.15	0.53
Approximate weights w_g (4.5.4)	0.13	0.59	0.18	0.44	0.13	0.62

Figure 4.2: Row 1 shows the exact weights w_g computed via Equation (4.5.2) using Monte Carlo simulation. These are the ideal weights, but they are not available in closed-form. Row 2 shows the approximate weights computed using Equation (4.5.4). As we can see, the weights are far from uniform, and the approximate weights are close to the exact weights.

terms should be zero), one might expect some terms $\left\|\frac{\partial \ell}{\partial \theta_g}\right\|$ to have a better random chance of being non-zero (for example, those of bigger dimensions). Thus for all parameters to be on equal footing, we would like to choose the weights w such that

$$E_{p_F}\left\|\frac{\partial \ell}{\partial \theta_g}\right\| = \text{constant} \times w_g, \tag{4.5.2}$$

where p_F is the fully factorized (independence) model. We will refer to these as the exact weights. These weights do not have a closed form expression, so we propose an approximation to these. It is simpler to compute in closed form $E_{p_F}\left\|\frac{\partial \ell}{\partial \theta_g}\right\|^2$, so we may use approximate weights

$$w_g \propto \sqrt{E_{p_F}\left\|\frac{\partial \ell}{\partial \theta_g}\right\|^2} \tag{4.5.3}$$

In the supplementary material, we show that the approximate weights (4.5.4) are

$$\begin{aligned} w_{st} &= \sigma_s \sigma_t \\ w_{sj} &= \sigma_s \sqrt{\sum_a p_a(1-p_a)} \\ w_{rj} &= \sqrt{\sum_a p_a(1-p_a) \sum_b q_b(1-q_b)} \end{aligned} \tag{4.5.4}$$

σ_s is the standard deviation of the continuous variable x_s. $p_a = Pr(y_r = a)$ and $q_b = Pr(y_j = b)$. For all 3 types of parameters, the weight has the form of $w_{uv} = \mathbf{tr}(\mathbf{cov}(z_u))\mathbf{tr}(\mathbf{cov}(z_v))$, where z represents a generic variable and $\mathbf{cov}(z)$ is the variance-covariance matrix of z.

We conducted a simulation study to show that calibration is needed. Consider a model with 4 independent variables: 2 continuous with variance 10 and 1, and 2 discrete variables with 10 and 2 levels.

There are 6 candidate edges in this model and from row 1 of Table 4.2 we can see the sizes of the gradients are different. In fact, the ratio of the largest gradient to the smallest gradient is greater than 4. The edges ρ_{11} and ρ_{12} involving the first continuous variable with variance 10 have large edge weights, than the corresponding edges, ρ_{21} and ρ_{22} involving the second continuous variable with variance 1. Similarly, the edges involving the first discrete variable with 10 levels are larger than the edges involving the second discrete variable with 2 levels. This reflects our intuition that larger variance and longer vectors will have larger norm.

Had the calibration weights been chosen via Equation 4.5.2, $w = \{w_g\}_g$ and the vector of gradients $\nabla\ell = \{\left\|\frac{\partial\ell}{\partial\theta_g}\right\|\}_g$ would have cosine similarity, $sim(u,v) = \frac{u^T v}{\|u\|\|v\|} = 1$. The approximate weights we used are from Equation (4.5.4) and have cosine similarity

$$sim(w, \nabla\ell) = .993,$$

which is extremely close to 1. Thus the calibration weights are effective in accounting for the size and variances of each edge group.

In the second simulation study, we used a model with 3 independent variables: one continuous, and 2 discrete variables with 2 and 4 levels. There are 3 candidate edges, and we computed the probability that a given edge would be the first allowed to enter the model using 3 different calibration schemes. From Table 4.1, we see that the uncalibrated regularizer would select the edge between the continuous variable and the 4 level discrete variable about 73% of the time. A perfect calibration scheme would select each edge 33% of the time. We see that the two proposed calibration schemes are an improvement over the uncalibrated regularizer.

The exact weights do not have a simple closed form expression, but they can be easily computed via Monte Carlo. This can be done by simulating independent Gaussians and multinomials with the appropriate marginal variance σ_s and marginal probabilities p_a, then approximating the expectation in (4.5.2) by an average. The computational cost of this procedure is negligible compared to fitting the mixed model, so the exact weights can also be used.

	ρ_{11}	ρ_{12}	ϕ_{12}
No Calibration $w_g = 1$	0.1350	0.7280	0.1370
Exact w_g (4.5.2)	0.3180	0.3310	0.3510
Approximate w_g (4.5.4)	0.2650	0.2650	0.4700

Table 4.1: Frequency an edge is the first selected by the group lasso regularizer. The group lasso with equal weights is highly unbalanced, as seen in row 1. The weighing scheme with the weights from (4.5.2) is very good, and selects the edges with probability close to the ideal $\frac{1}{3}$. The approximate weighing scheme of (4.5.4) is an improvement over not calibrating; however, not as good as the weights from (4.5.2).

4.6 Model Selection Consistency

In this section, we study the model selection consistency, whether the correct edge set is selected and the parameter estimates are close to the truth, of the pseudolikelihood and maximum likelihood estimators. Consistency can be established using the framework first developed in Ravikumar et al. (2010) and later extended to general m-estimators by Lee et al. (2013b). Instead of stating the full results and proofs, we will illustrate the type of theorems that can be shown and defer the rigorous statements to the Supplementary Material.

First, we define some notation. Recall that Θ is the vector of parameters being estimated $\{\beta_{ss}, \beta_{st}, \alpha_s, \phi_{rj}, \rho_{sj}\}$, $\Theta^\star$ be the true parameters that estimated the model, and $Q = \nabla^2 \ell(\Theta^\star)$. Both maximum likelihood and pseudolikelihood estimation procedures can be written as a convex optimization problem of the form

$$\text{minimize } \ell(\Theta) + \lambda \sum_{g \in G} \|\Theta_g\|_2 \tag{4.6.1}$$

where $\ell(\theta) = \{\ell_{ML}, \ell_{PL}\}$ is one of the two log-likelihoods. The regularizer

$$\sum_{g \in G} \|\Theta_g\| = \lambda \left(\sum_{s=1}^{p} \sum_{t=1}^{s-1} |\beta_{st}| + \sum_{s=1}^{p} \sum_{j=1}^{q} \|\rho_{sj}\|_2 + \sum_{j=1}^{q} \sum_{r=1}^{j-1} \|\phi_{rj}\|_F \right).$$

The set G indexes the edges β_{st}, ρ_{sj}, and ϕ_{rj}, and Θ_g is one of the three types of edges. Let A and I represent the active and inactive groups in Θ, so $\Theta_g^\star \neq 0$ for any $g \in A$ and $\Theta_g^\star = 0$ for any $g \in I$.

Let $\hat{\Theta}$ be the minimizer to Equation (4.6.1). Then $\hat{\Theta}$ satisfies,

1. $\left\|\hat{\Theta} - \Theta^\star\right\|_2 \leq C\sqrt{\frac{|A| \log |G|}{n}}$
2. $\hat{\Theta}_g = 0$ for $g \in I$.

The exact statement of the theorem is given in the Supplementary Material.

4.7 Optimization Algorithms

In this section, we discuss two algorithms for solving (4.4.2): the proximal gradient and the proximal newton methods. This is a convex optimization problem that decomposes into the form $f(x) + g(x)$, where f is smooth and convex and g is convex but possibly non-smooth. In our case f is the negative log-likelihood or negative log-pseudolikelihood and g are the group sparsity penalties.

Block coordinate descent is a frequently used method when the non-smooth function g is the l_1 or group l_1. It is especially easy to apply when the function f is quadratic, since each block coordinate update can be solved in closed form for many different non-smooth g (Friedman et al., 2007). The smooth f in our particular case is not quadratic, so each block update cannot be solved in closed form. However in certain problems (sparse inverse covariance), the update can be approximately solved by using an appropriate inner optimization routine (Friedman et al., 2008b).

4.7.1 Proximal Gradient

Problems of this form are well-suited for the proximal gradient and accelerated proximal gradient algorithms as long as the proximal operator of g can be computed (Combettes and Pesquet, 2011; Beck and Teboulle, 2010)

$$prox_t(x) = \arg\min_u \frac{1}{2t} \|x - u\|^2 + g(u) \tag{4.7.1}$$

For the sum of l_2 group sparsity penalties considered, the proximal operator takes the familiar form of soft-thresholding and group soft-thresholding (Bach et al., 2011). Since the groups are non-overlapping, the proximal operator simplifies to scalar soft-thresholding for β_{st} and group soft-thresholding for ρ_{sj} and ϕ_{rj}.

The class of proximal gradient and accelerated proximal gradient algorithms is directly applicable to our problem. These algorithms work by solving a first-order model at the

current iterate x_k

$$\arg\min_u \; f(x_k) + \nabla f(x_k)^T(u - x_k) + \frac{1}{2t}\|u - x_k\|^2 + g(u) \tag{4.7.2}$$

$$= \arg\min_u \; \frac{1}{2t}\|u - (x_k - t\nabla f(x_k))\|^2 + g(u) \tag{4.7.3}$$

$$= prox_t(x_k - t\nabla f(x_k)) \tag{4.7.4}$$

The proximal gradient iteration is given by $x_{k+1} = prox_t\left(x_k - t\nabla f(x_k)\right)$ where t is determined by line search. The theoretical convergence rates and properties of the proximal gradient algorithm and its accelerated variants are well-established (Beck and Teboulle, 2010). The accelerated proximal gradient method achieves linear convergence rate of $O(c^k)$ when the objective is strongly convex and the sublinear rate $O(1/k^2)$ for non-strongly convex problems.

The TFOCS framework (Becker et al., 2011) is a package that allows us to experiment with 6 different variants of the accelerated proximal gradient algorithm. The TFOCS authors found that the Auslender-Teboulle algorithm exhibited less oscillatory behavior, and proximal gradient experiments in the next section were done using the Auslender-Teboulle implementation in TFOCS.

4.7.2 Proximal Newton Algorithms

The class of proximal Newton algorithms is a 2nd order analog of the proximal gradient algorithms with a quadratic convergence rate (Lee et al., 2012; Schmidt, 2010; Schmidt et al., 2011). It attempts to incorporate 2nd order information about the smooth function f into the model function. At each iteration, it minimizes a quadratic model centered at x_k

$$\arg\min_u \; f(x_k) + \nabla f(x_k)^T(u - x_k) + \frac{1}{2t}(u - x_k)^T H(u - x_k) + g(u) \tag{4.7.5}$$

$$= \arg\min_u \; \frac{1}{2t}\left(u - x_k + tH^{-1}\nabla f(x_k)\right)^T H\left(u - x_k + tH^{-1}\nabla f(x_k)\right) + g(u) \tag{4.7.6}$$

$$= \arg\min_u \; \frac{1}{2t}\left\|u - \left(x_k - tH^{-1}\nabla f(x_k)\right)\right\|_H^2 + g(u) \tag{4.7.7}$$

$$:= Hprox_t\left(x_k - tH^{-1}\nabla f(x_k)\right) \text{ where } H = \nabla^2 f(x_k) \tag{4.7.8}$$

The $Hprox$ operator is analogous to the proximal operator, but in the $\|\cdot\|_H$-norm. It simplifies to the proximal operator if $H = I$, but in the general case of positive definite H

Algorithm 3 Proximal Newton

repeat

Solve subproblem $p_k = Hprox_t\left(x_k - tH_k^{-1}\nabla f(x_k)\right) - x_k$ using TFOCS.

Find t to satisfy Armijo line search condition with parameter α

$$f(x_k + tp_k) + g(x_k + tp_k) \leq f(x_k) + g(x_k) - \frac{t\alpha}{2}\|p_k\|^2$$

Set $x_{k+1} = x_k + tp_k$

$k = k + 1$

until $\frac{\|x_k - x_{k+1}\|}{\|x_k\|} < tol$

there is no closed-form solution for many common non-smooth $g(x)$ (including l_1 and group l_1). However if the proximal operator of g is available, each of these sub-problems can be solved efficiently with proximal gradient. In the case of separable g, coordinate descent is also applicable. Fast methods for solving the subproblem $Hprox_t(x_k - tH^{-1}\nabla f(x_k))$ include coordinate descent methods, proximal gradient methods, or Barzilai-Borwein (Friedman et al., 2007; Combettes and Pesquet, 2011; Beck and Teboulle, 2010; Wright et al., 2009). The proximal Newton framework allows us to bootstrap many previously developed solvers to the case of arbitrary loss function f.

Theoretical analysis in Lee et al. (2012) suggests that proximal Newton methods generally require fewer outer iterations (evaluations of $Hprox$) than first-order methods while providing higher accuracy because they incorporate 2nd order information. We have confirmed empirically that the proximal Newton methods are faster when n is very large or the gradient is expensive to compute (e.g. maximum likelihood estimation). Since the objective is quadratic, coordinate descent is also applicable to the subproblems. The hessian matrix H can be replaced by a quasi-newton approximation such as BFGS/L-BFGS/SR1. In our implementation, we use the `PNOPT` implementation (Lee et al., 2012).

4.7.3 Path Algorithm

Frequently in machine learning and statistics, the regularization parameter λ is heavily dependent on the dataset. λ is generally chosen via cross-validation or holdout set performance, so it is convenient to provide solutions over an interval of $[\lambda_{min}, \lambda_{max}]$. We start the algorithm at $\lambda_1 = \lambda_{max}$ and solve, using the previous solution as warm start, for $\lambda_2 > \ldots > \lambda_{min}$. We find that this reduces the cost of fitting an entire path of solutions

(See Figure 4.5). λ_{max} can be chosen as the smallest value such that all parameters are 0 by using the KKT equations (Friedman et al., 2007).

4.8 Conditional Model

In addition to the variables we would like to model, there are often additional features or covariates that affect the dependence structure of the variables. For example in genomic data, in addition to expression values, we have attributes associated to each subject such as gender, age and ethnicity. These additional attributes affect the dependence of the expression values, so we can build a conditional model that uses the additional attributes as features. In this section, we show how to augment the pairwise mixed model with features.

Conditional models only model the conditional distribution $p(z|f)$, as opposed to the joint distribution $p(z, f)$, where z are the variables of interest to the prediction task and f are features. These models are frequently used in practice Lafferty et al. (2001).

In addition to observing x and y, we observe features f and we build a graphical model for the conditional distribution $p(x, y|f)$. Consider a full pairwise model $p(x, y, f)$ of the form (4.2.1). We then choose to only model the joint distribution over only the variables x and y to give us $p(x, y|f)$ which is of the form

$$\begin{aligned} p(x,y|f;\Theta) =& \frac{1}{Z(\Theta|f)} \exp\left(\sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}x_s x_t + \sum_{s=1}^{p}\alpha_s x_s + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j)x_s \right. \\ &\left. + \sum_{j=1}^{q}\sum_{r=1}^{j}\phi_{rj}(y_r,y_j) + \sum_{l=1}^{F}\sum_{s=1}^{p}\gamma_{ls}x_s f_l + \sum_{l=1}^{F}\sum_{r=1}^{q}\eta_{lr}(y_r)f_l\right) \end{aligned} \tag{4.8.1}$$

We can also consider a more general model where each pairwise edge potential depends on the features

$$\begin{aligned} p(x,y|f;\Theta) = \frac{1}{Z(\Theta|f)} \exp\left(\sum_{s=1}^{p}\sum_{t=1}^{p} -\frac{1}{2}\beta_{st}(f)x_s x_t + \sum_{s=1}^{p}\alpha_s(f)x_s \right. \\ \left. + \sum_{s=1}^{p}\sum_{j=1}^{q}\rho_{sj}(y_j,f)x_s + \sum_{j=1}^{q}\sum_{r=1}^{j}\phi_{rj}(y_r,y_j,f)\right) \end{aligned} \tag{4.8.2}$$

(4.8.1) is a special case of this where only the node potentials depend on features and the pairwise potentials are independent of feature values. The specific parametrized form we

consider is $\phi_{rj}(y_r, y_j, f) \equiv \phi_{rj}(y_r, y_j)$ for $r \neq j$, $\rho_{sj}(y_j, f) \equiv \rho_{sj}(y_j)$, and $\beta_{st}(f) = \beta_{st}$. The node potentials depend linearly on the feature values, $\alpha_s(f) = \alpha_s + \sum_{l=1}^{F} \gamma_{ls} x_s f_l$, and $\phi_{rr}(y_r, y_r, f) = \phi_{rr}(y_r, y_r) + \sum_l \eta_{lr}(y_r)$.

4.9 Experimental Results

We present experimental results on synthetic data, survey data and on a conditional model.

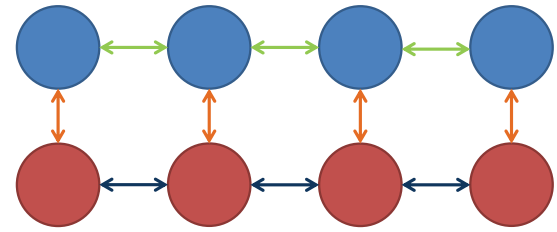
4.9.1 Synthetic Experiments

In the synthetic experiment, the training points are sampled from a true model with 10 continuous variables and 10 binary variables. The edge structure is shown in Figure 4.3a. λ is chosen proportional to $\sqrt{\frac{\log(p+q)}{n}}$ as suggested by the theoretical results in Section 4.6. We experimented with 3 values $\lambda = \{1, 5, 10\}\sqrt{\frac{\log(p+q)}{n}}$ and chose $\lambda = 5\sqrt{\frac{\log(p+q)}{n}}$ so that the true edge set was recovered by the algorithm for the sample size $n = 2000$. We see from the experimental results that recovery of the correct edge set undergoes a sharp phase transition, as expected. With $n = 1000$ samples, the pseudolikelihood is recovering the correct edge set with probability nearly 1. The maximum likelihood was performed using an exact evaluation of the gradient and log-partition. The poor performance of the maximum likelihood estimator is explained by the maximum likelihood objective violating the irrepresentable condition; a similar example is discussed in (Ravikumar et al., 2010, Section 3.1.1), where the maximum likelihood is not irrepresentable, yet the neighborhood selection procedure is. The phase transition experiments were done using the proximal Newton algorithm discussed in Section 4.7.2.

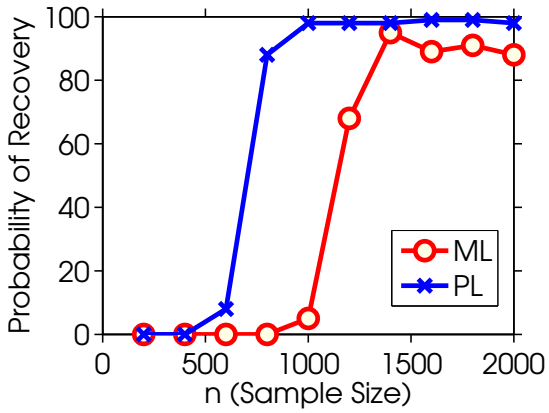
We also run the proximal Newton algorithm for a sequence of instances with $p = q = 10, 50, 100, 500, 1000$ and $n = 500$. The largest instance has 2000 variables and takes 12.5 hours to complete. The timing results are summarized in Figure 4.4.

4.9.2 Survey Experiments

The census survey dataset we consider consists of 11 variables, of which 2 are continuous and 9 are discrete: age (continuous), log-wage (continuous), year(7 states), sex(2 states),marital status (5 states), race(4 states), education level (5 states), geographic region(9 states), job class (2 states), health (2 states), and health insurance (2 states). The dataset was assembled



(a)



(b)

Figure 4.3: *Figure 4.3a shows the graph used in the synthetic experiments for $p = q = 4$; the experiment actually used p=10 and q=10. Blue nodes are continuous variables, red nodes are binary variables and the orange, green and dark blue lines represent the 3 types of edges. Figure 4.3b is a plot of the probability of correct edge recovery, meaning every true edge is selected and no non-edge is selected, at a given sample size using Maximum Likelihood and Pseudolikelihood. Results are averaged over 100 trials.*

$p+q$	Time per Iteration (sec)	Total Time (min)	Number of Iterations
20	.13	.003	13
100	4.39	1.32	18
200	18.44	6.45	21
1000	245.34	139	34
2000	1025.6	752	44

Figure 4.4: Timing experiments for various instances of the graph in Figure 4.3a. The number of variables range from 20 to 2000 with $n = 500$.

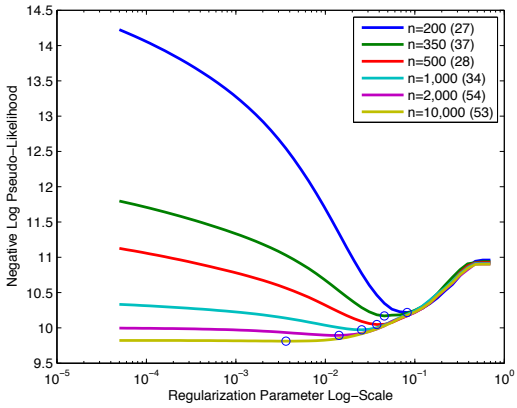


Figure 4.5: *Model selection under different training set sizes. Circle denotes the lowest test set negative log pseudolikelihood and the number in parentheses is the number of edges in that model at the lowest test negative log pseudolikelihood. The saturated model has* 55 *edges.*

by Steve Miller of OpenBI.com from the March 2011 Supplement to Current Population Survey data. All the evaluations are done using a holdout test set of size $100,000$ for the survey experiments. The regularization parameter λ is varied over the interval $[5 \times 10^{-5}, 0.7]$ at 50 points equispaced on log-scale for all experiments. In practice, λ can be chosen to minimize the holdout log pseudolikelihood.

Model Selection

In Figure 4.5, we study the model selection performance of learning a graphical model over the 11 variables under different training samples sizes. We see that as the sample size increases, the optimal model is increasingly dense, and less regularization is needed.

Comparing against Separate Regressions

A sensible baseline method to compare against is a separate regression algorithm. This algorithm fits a linear Gaussian or (multiclass) logistic regression of each variable conditioned on the rest. We can evaluate the performance of the pseudolikelihood by evaluating $-\log p(x_s|x_{\setminus s}, y)$ for linear regression and $-\log p(y_r|y_{\setminus r}, x)$ for (multiclass) logistic regression. Since regression is directly optimizing this loss function, it is expected to do better. The pseudolikelihood objective is similar, but has half the number of parameters as the separate regressions since the coefficients are shared between two of the conditional likelihoods. From Figures 4.6 and 4.7, we can see that the pseudolikelihood performs very similarly to

the separate regressions and sometimes even outperforms regression. The benefit of the pseudolikelihood is that we have learned parameters of the joint distribution $p(x, y)$ and not just of the conditionals $p(x_s|y, x_{\setminus s})$. On the test dataset, we can compute quantities such as conditionals over arbitrary sets of variables $p(y_A, x_B|y_{A^C}, x_{B^C})$ and marginals $p(x_A, y_B)$ (Koller and Friedman, 2009). This would not be possible using the separate regressions.

Conditional Model

Using the conditional model (4.8.1), we model only the 3 variables logwage, education(5) and jobclass(2). The other 8 variables are only used as features. The conditional model is then trained using the pseudolikelihood. We compare against the generative model that learns a joint distribution on all 11 variables. From Figure 4.8, we see that the conditional model outperforms the generative model, except at small sample sizes. This is expected since the conditional distribution models less variables. At very small sample sizes and small λ, the generative model outperforms the conditional model. This is likely because generative models converge faster (with less samples) than discriminative models to its optimum.

Maximum Likelihood vs Pseudolikelihood

The maximum likelihood estimates are computable for very small models such as the conditional model previously studied. The pseudolikelihood was originally motivated as an approximation to the likelihood that is computationally tractable. We compare the maximum likelihood and maximum pseudolikelihood on two different evaluation criteria: the negative log likelihood and negative log pseudolikelihood. In Figure 4.9, we find that the pseudolikelihood outperforms maximum likelihood under both the negative log likelihood and negative log pseudolikelihood. We would expect that the pseudolikelihood trained model does better on the pseudolikelihood evaluation and maximum likelihood trained model does better on the likelihood evaluation. However, we found that the pseudolikelihood trained model outperformed the maximum likelihood trained model on both evaluation criteria. Although asymptotic theory suggests that maximum likelihood is more efficient than the pseudolikelihood, this analysis is applicable because of the finite sample regime and misspecified model. See Liang and Jordan (2008) for asymptotic analysis of pseudolikelihood and maximum likelihood under a well-specified model. We also observed the pseudolikelihood slightly outperforming the maximum likelihood in the synthetic experiment of Figure

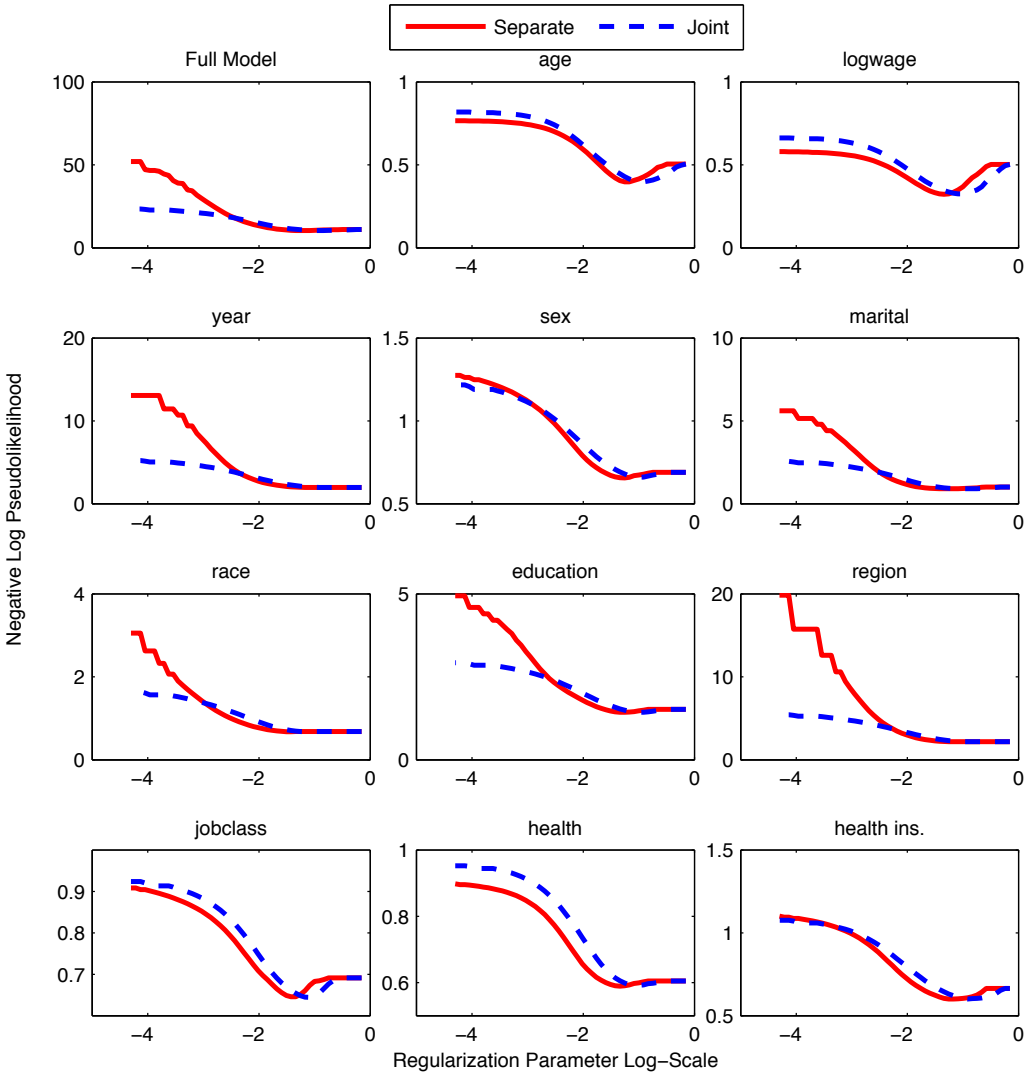


Figure 4.6: *Separate Regression vs Pseudolikelihood* $n = 100$. *y-axis is the appropriate regression loss for the response variable. For low levels of regularization and at small training sizes, the pseudolikelihood seems to overfit less; this may be due to a global regularization effect from fitting the joint distribution as opposed to separate regressions.*

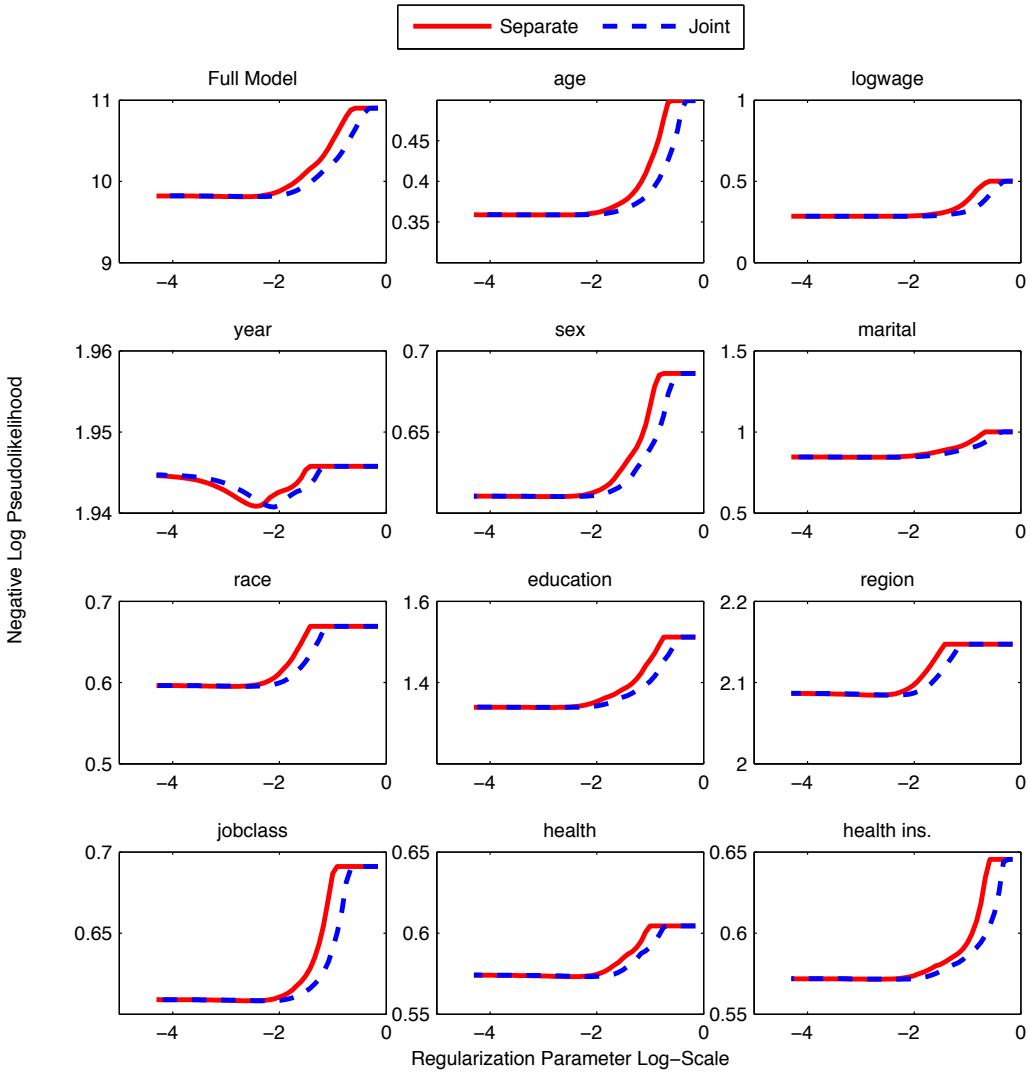


Figure 4.7: *Separate Regression vs Pseudolikelihood* $n = 10,000$. *y-axis is the appropriate regression loss for the response variable. At large sample sizes, separate regressions and pseudolikelihood perform very similarly. This is expected since this is nearing the asymptotic regime.*

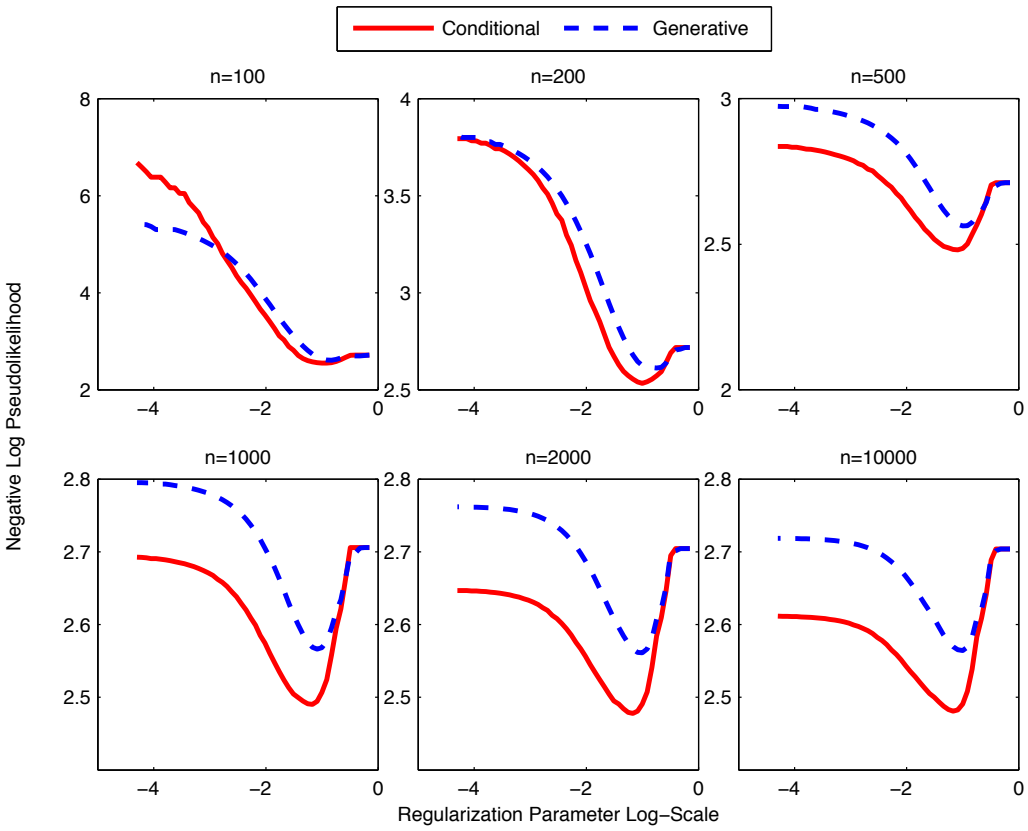


Figure 4.8: *Conditional Model vs Generative Model at various sample sizes. y-axis is test set performance is evaluated on negative log pseudolikelihood of the conditional model. The conditional model outperforms the full generative model at except the smallest sample size* $n = 100$.

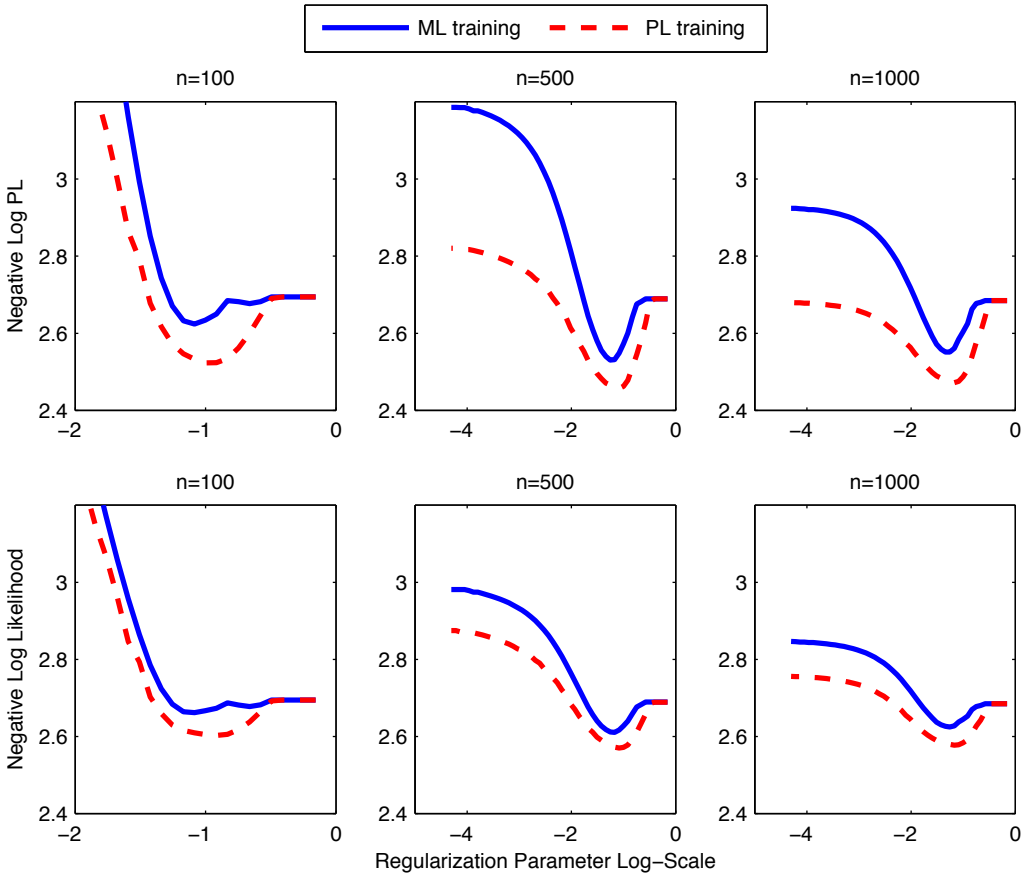


Figure 4.9: *Maximum Likelihood vs Pseudolikelihood. y-axis for top row is the negative log pseudolikelihood. y-axis for bottom row is the negative log likelihood. Pseudolikelihood outperforms maximum likelihood across all the experiments.*

4.3b.

4.10 Conclusion

This work proposes a new pairwise mixed graphical model, which combines the Gaussian graphical model and discrete graphical model. Due to the introduction of discrete variables, the maximum likelihood estimator is computationally intractable, so we investigated the pseudolikelihood estimator. To learn the structure of this model, we use the appropriate group sparsity penalties with a calibrated weighing scheme. Model selection consistency results are shown for the mixed model using the maximum likelihood and pseudolikelihood estimators. The extension to a conditonal model is discussed, since these are frequently

used in practice.

We proposed two efficient algorithms for the purpose of estimating the parameters of this model, the proximal Newton and the proximal gradient algorithms. The proximal Newton algorithm is shown to scale to graphical models with 2000 variables on a standard desktop. The model is evaluated on synthetic and the current population survey data, which demonstrates the pseudolikelihood performs well compared to maximum likelihood and nodewise regression.

For future work, it would be interesting to incorporate other discrete variables such as poisson or binomial variables and non-Gaussian continuous variables. This would broaden the scope of applications that mixed models could be used for. Our work is a first step in that direction.

Supplementary Materials

4.10.1 Proof of Convexity

Proposition 4.3.1. *The negative log pseudolikelihood in (4.3.3) is jointly convex in all the parameters* $\{\beta_{ss}, \beta_{st}, \alpha_s, \phi_{rj}, \rho_{sj}\}$ *over the region* $\beta_{ss} > 0$.

Proof. To verify the convexity of $\tilde{\ell}(\Theta|x, y)$, it suffices to check that each term is convex. $-\log p(y_r|y_{\backslash r,}, x; \Theta)$ is jointly convex in ρ and ϕ since it is a multiclass logistic regression. We now check that $-\log p(x_s|x_{\backslash s}, y; \Theta)$ is convex. $-\frac{1}{2}\log \beta_{ss}$ is a convex function. To establish that

$$\frac{\beta_{ss}}{2}\left(\frac{\alpha_s}{\beta_{ss}} + \sum_j \frac{\rho_{sj}(y_j)}{\beta_{ss}} - \sum_{t\neq s}\frac{\beta_{st}}{\beta_{ss}}x_t - x_s\right)^2$$

is convex, we use the fact that $f(u, v) = \frac{v}{2}(\frac{u}{v} - c)^2$ is convex. Let $v = \beta_{ss}$, $u = \alpha_s + \sum_j \rho_{sj}(y_j) - \sum_{t\neq s}\beta_{st}x_t$, and $c = x_s$. Notice that x_s, α_s, y_j, and x_t are fixed quantities and u is affinely related to β_{st} and ρ_{sj}. A convex function composed with an affine map is still convex, thus $\frac{\beta_{ss}}{2}\left(\frac{\alpha_s}{\beta_{ss}} + \sum_j \frac{\rho_{sj}(y_j)}{\beta_{ss}} - \sum_{t\neq s}\frac{\beta_{st}}{\beta_{ss}}x_t - x_s\right)^2$ is convex.

To finish the proof, we verify that $f(u, v) = \frac{v}{2}(\frac{u}{v} - c)^2 = \frac{1}{2}\frac{(u-cv)^2}{v}$ is convex over $v > 0$. The epigraph of a convex function is a convex set iff the function is convex. Thus we establish that the set $C = \{(u, v, t)|\frac{1}{2}\frac{(u-cv)^2}{v} \leq t, v > 0\}$ is convex. Let $A = \begin{bmatrix} v & u - cv \\ u - cv & t \end{bmatrix}$. The Schur complement criterion of positive definiteness says $A \succ 0$ iff $v > 0$ and $t > \frac{(u-cv)^2}{v}$. The condition $A \succ 0$ is a linear matrix inequality and thus convex in the entries of A. The entries of A are linearly related to u and v, so $A \succ 0$ is also convex in u and v. Therefore $v > 0$ and $t > \frac{(u-cv)^2}{v}$ is a convex set. □

4.10.2 Sampling From The Joint Distribution

In this section we discuss how to draw samples $(x, y) \sim p(x, y)$. Using the property that $p(x, y) = p(y)p(x|y)$, we see that if $y \sim p(y)$ and $x \sim p(x|y)$ then $(x, y) \sim p(x, y)$. We have

that

$$p(y) \propto \exp\left(\sum_{r,j} \phi_{rj}(y_r, y_j) + \frac{1}{2}\rho(y)^T B^{-1}\rho(y)\right) \tag{4.10.1}$$

$$(\rho(y))_s = \sum_j \rho_{sj}(y_j) \tag{4.10.2}$$

$$p(x|y) = No(B^{-1}(\alpha + \rho(y)), B^{-1}) \tag{4.10.3}$$

The difficult part is to sample $y \sim p(y)$ since this involves the partition function of the discrete MRF. This can be done with MCMC for larger models and junction tree algorithm or exact sampling for small models.

4.10.3 Maximum Likelihood

The difficulty in MLE is that in each gradient step we have to compute $\hat{T}(x,y) - E_{p(\Theta)}[T(x,y)]$, the difference between the empirical sufficient statistic $\hat{T}(x,y)$ and the expected sufficient statistic. In both continuous and discrete graphical models the computationally expensive step is evaluating $E_{p(\Theta)}[T(x,y)]$. In discrete problems, this involves a sum over the discrete state space and in continuous problem, this requires matrix inversion. For both discrete and continuous models, there has been much work on addressing these difficulties. For discrete models, the junction tree algorithm is an exact method for evaluating marginals and is suitable for models with low tree width. Variational methods such as belief propagation and tree reweighted belief propagation work by optimizing a surrogate likelihood function by approximating the partition function $Z(\Theta)$ by a tractable surrogate $\widetilde{Z}(\Theta)$ Wainwright and Jordan (2008). In the case of a large discrete state space, these methods can be used to approximate $p(y)$ and do approximate maximum likelihood estimation for the discrete model. Approximate maximum likelihood estimation can also be done via Monte Carlo estimates of the gradients $\hat{T}(x,y) - E_{p(\Theta)}(T(x,y))$. For continuous Gaussian graphical models, efficient algorithms based on block coordinate descent Friedman et al. (2008b); Banerjee et al. (2008) have been developed, that do not require matrix inversion.

The joint distribution and loglikelihood are:

$$p(x, y; \Theta) = \exp\left(-\frac{1}{2}x^T Bx + (\alpha + \rho(y))^T x + \sum_{(r,j)} \phi_{rj}(y_r, y_j)\right)/Z(\Theta)$$

$$\ell(\Theta) = \left(\frac{1}{2}x^T Bx - (\alpha + \rho(y))^T x - \sum_{(r,j)} \phi_{rj}(y_r, y_j)\right) + \log\left(\sum_{y'} \int dx \exp\left(-\frac{1}{2}x^T Bx + (\alpha + \rho(y'))^T x\right) \exp\left(\sum_{(r,j)} \phi_{rj}(y'_r, y'_j)\right)\right)$$

The derivative is

$$\begin{aligned}
\frac{\partial \ell}{\partial B} &= \frac{1}{2}xx^T + \frac{\int dx\left(\sum_{y'} -\frac{1}{2}xx^T \exp\left(-\frac{1}{2}x^T Bx + (\alpha + \rho(y))^T x + \sum_{(r,j)} \phi_{rj}(y'_r, y'_j)\right)\right)}{Z(\Theta)} \\
&= \frac{1}{2}xx^T + \int \sum_{y'} \left(-\frac{1}{2}xx^T p(x, y'; \Theta)\right) \\
&= \frac{1}{2}xx^T + \sum_{y'} \int -\frac{1}{2}xx^T p(x|y'; \Theta) p(y') \\
&= \frac{1}{2}xx^T + \sum_{y'} \int -\frac{1}{2}\left(B^{-1} + B^{-1}(\alpha + \rho(y'))(\alpha + \rho(y')^T)B^{-1}\right) p(y')
\end{aligned}$$

The primary cost is to compute B^{-1} and the sum over the discrete states y.
The computation for the derivatives of $\ell(\Theta)$ with respect to ρ_{sj} and ϕ_{rj} are similar.

$$\begin{aligned}
\frac{\partial \ell}{\phi_{rj}(a, b)} &= -1(y_r = a, y_j = b) + \sum_{y'} \int dx 1(y'_r = a, y'_j = b) p(x, y'; \Theta) \\
&= -1(y_r = a, y_j = b) + \sum_{y'} 1(y'_r = a, y'_j = b) p(y')
\end{aligned}$$

The gradient requires summing over all discrete states.

Similarly for $\rho_{sj}(a)$:

$$\frac{\partial \ell}{\rho_{sj}(a)} = -1(y_j = a)x_s + \sum_{y'} \int dx (1(y'_j = a)x_s) p(x', y'; \Theta)$$
$$= -1(y_j = a)x_s + \int dx \sum_{y'_{\setminus j}} x_s p(x|y'_{\setminus j}, y'_j = a) p(y'_{\setminus j}, y'_j = a)$$

MLE estimation requires summing over the discrete states to compute the expected sufficient statistics. This may be approximated using using samples $(x, y) \sim p(x, y; \Theta)$. The method in the previous section shows that sampling is efficient if $y \sim p(y)$ is efficient. This allows us to use MCMC methods developed for discrete MRF's such as Gibbs sampling.

4.10.4 Choosing the Weights

We first show how to compute w_{sj}. The gradient of the pseudo-likelihood with respect to a parameter $\rho_{sj}(a)$ is given below

$$\frac{\partial \tilde{\ell}}{\partial \rho_{sj}(a)} = \sum_{i=1}^{n} -2 \times \mathbb{1}\left[y_j^i = a\right] x_s^i + E_{p_F}(\mathbb{1}[y_j = a]x_s | y^i_{\setminus j}, x^i) + E_{p_F}(\mathbb{1}[y_j = a]x_s | x^i_{\setminus s}, y^i)$$
$$= \sum_{i=1}^{n} -2 \times \mathbb{1}\left[y_j^i = a\right] x_s^i + x_s^i p(y_j = a) + \mathbb{1}\left[y_j^i = a\right] \mu_s$$
$$= \sum_{i=1}^{n} \mathbb{1}\left[y_j^i = a\right] \left(\hat{\mu}_s - x_s^i\right) + x_s^i \left(\hat{p}(y_j = a) - \mathbb{1}\left[y_j^i = a\right]\right)$$
$$= \sum_{i=1}^{n} \left(\mathbb{1}\left[y_j^i = a\right] - \hat{p}(y_j = a)\right) \left(\hat{\mu}_s - x_s^i\right) + \left(x_s^i - \hat{\mu}_s\right) \left(\hat{p}(y_j = a) - \mathbb{1}\left[y_j^i = a\right]\right) \tag{4.10.4}$$
$$= \sum_{i=1}^{n} 2 \left(\mathbb{1}\left[y_j^i = a\right] - \hat{p}(y_j = a)\right) \left(\hat{\mu}_s - x_s^i\right) \tag{4.10.5}$$

Since the subgradient condition includes a variable if $\left\|\frac{\partial \tilde{\ell}}{\partial \rho_{sj}}\right\| > \lambda$, we compute $E\left\|\frac{\partial \tilde{\ell}}{\partial \rho_{sj}}\right\|^2$. By independence,

$$E_{p_F}\left(\left\|\sum_{i=1}^{n} 2\left(\mathbb{1}\left[y_j^i = a\right] - \hat{p}(y_j = a)\right)\left(\hat{\mu}_s - x_s^i\right)\right\|^2\right) \tag{4.10.6}$$

$$= 4nE_{p_F}\left(\left\|\mathbb{1}\left[y_j^i = a\right] - \hat{p}(y_j = a)\right\|^2\right) E_{p_F}\left(\left\|\hat{\mu}_s - x_s^i\right\|^2\right) \tag{4.10.7}$$

$$= 4(n-1)p(y_j = a)(1 - p(y_j = a))\sigma_s^2 \tag{4.10.8}$$

The last line is an equality if we replace the sample means $\hat{p}$ and $\hat{\mu}$ with the true values p and μ. Thus for the entire vector ρ_{sj} we have $E_{p_F}\left\|\frac{\partial \tilde{\ell}}{\partial \rho_{sj}}\right\|^2 = 4(n-1)\left(\sum_a p(y_j = a)(1 - p(y_j = a))\right)\sigma_s^2$. If we let the vector z be the indicator vector of the categorical variable y_j, and let the vector $p = p(y_j = a)$, then $E_{p_F}\left\|\frac{\partial \tilde{\ell}}{\partial \rho_{sj}}\right\|^2 = 4(n-1)\sum_a p_a(1-p_a)\sigma^2 = 4(n-1)\mathbf{tr}(\mathbf{cov}(z))\mathbf{var}(x)$ and $w_{sj} = \sqrt{\sum_a p_a(1-p_a)\sigma_s^2}$.

We repeat the computation for β_{st}.

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta_{st}} &= \sum_{i=1}^{n} -2x_s^i x_t + E_{p_F}(x_s^i x_t^i | x_{\setminus s}, y) + E_{p_F}(x_s^i x_t^i | x_{\setminus t}, y) \\
&= \sum_{i=1}^{n} -2x_s^i x_t^i + \hat{\mu}_s x_t^i + \hat{\mu}_t x_s^i \\
&= \sum_{i=1}^{n} x_t^i(\hat{\mu_s} - x_s^i) + x_s^i(\hat{\mu}_t - x_t^i) \\
&= \sum_{i=1}^{n} (x_t^i - \hat{\mu}_t)(\hat{\mu_s} - x_s^i) + (x_s^i - \hat{\mu_s})(\hat{\mu}_t - x_t^i) \\
&= \sum_{i=1}^{n} 2(x_t^i - \hat{\mu}_t)(\hat{\mu_s} - x_s^i)
\end{aligned}$$

Thus

$$\begin{aligned}
&E\left(\left\|\sum_{i=1}^{n} 2(x_t^i - \hat{\mu}_t)(\hat{\mu_s} - x_s^i)\right\|^2\right) \\
&= 4nE_{p_F}\left\|x_t - \hat{\mu_t}\right\|^2 E_{p_F}\left\|x_s - \hat{\mu}_s\right\|^2 \\
&= 4(n-1)\sigma_s^2\sigma_t^2
\end{aligned}$$

Thus $E_{p_F}\left\|\frac{\partial \ell}{\partial \beta_{st}}\right\|^2 = 4(n-1)\sigma_s^2\sigma_t^2$ and taking square-roots gives us $w_{st} = \sigma_s\sigma_t$.
We repeat the same computation for ϕ_{rj}. Let $p_a = Pr(y_r = a)$ and $q_b = Pr(y_j = b)$.

$$\begin{aligned}
\frac{\partial \tilde{\ell}}{\partial \phi_{rj}(a,b)} &= \sum_{i=1}^{n} -\mathbb{1}\left[y_r^i = a\right]\mathbb{1}\left[y_j^i = b\right] + E\left(\mathbb{1}[y_r = a]\mathbb{1}[y_j = b]|y_{\setminus r}, x\right) \\
&\quad + E\left(\mathbb{1}[y_r = a]\mathbb{1}[y_j = b]|y_{\setminus j}, x\right) \\
&= \sum_{i=1}^{n} -\mathbb{1}\left[y_r^i = a\right]\mathbb{1}\left[y_j^i = b\right] + \hat{p}_a\mathbb{1}\left[y_j^i = b\right] + \hat{q}_b\mathbb{1}\left[y_r^i = a\right] \\
&= \sum_{i=1}^{n} \mathbb{1}\left[y_j^i = b\right](\hat{p}_a - \mathbb{1}\left[y_r^i = a\right]) + \mathbb{1}\left[y_r^i = a\right](\hat{q}_b - \mathbb{1}\left[y_j^i = b\right]) \\
&= \sum_{i=1}^{n} (\mathbb{1}\left[y_j^i = b\right] - \hat{q}_b)(\hat{p}_a - \mathbb{1}\left[y_r^i = a\right]) + (\mathbb{1}\left[y_r^i = a\right] - \hat{p}_a)(\hat{q}_b - \mathbb{1}\left[y_j^i = b\right]) \\
&= \sum_{i=1}^{n} 2(\mathbb{1}\left[y_j^i = b\right] - \hat{q}_b)(\hat{p}_a - \mathbb{1}\left[y_r^i = a\right])
\end{aligned}$$

Thus we compute

$$\begin{aligned}
E_{p_F}\left\|\frac{\partial \tilde{\ell}}{\partial \phi_{rj}(a,b)}\right\|^2 &= E\left(\left\|\sum_{i=1}^{n} 2(\mathbb{1}\left[y_j^i = b\right] - \hat{q}_b)(\hat{p}_a - \mathbb{1}\left[y_r^i = a\right])\right\|^2\right) \\
&= 4nE_{p_F}\left\|\hat{q}_b - \mathbb{1}[y_j = b]\right\|^2 E_{p_F}\left\|\hat{p}_a - \mathbb{1}[y_r = a]\right\|^2 \\
&= 4(n-1)q_b(1-q_b)p_a(1-p_a)
\end{aligned}$$

From this, we see that $E_{p_F}\left\|\frac{\partial \tilde{\ell}}{\partial \phi_{rj}}\right\|^2 = \sum_{a=1}^{L_r}\sum_{b=1}^{L_j} 4(n-1)q_b(1-q_b)p_a(1-p_a)$ and $w_{rj} = \sqrt{\sum_{a=1}^{L_r}\sum_{b=1}^{L_j} q_b(1-q_b)p_a(1-p_a)}$.

4.10.5 Model Selection Consistency

One of the difficulties in establishing consistency results for the problem in Equation (4.6.1) is due to the non-identifiability of the parameters. $\ell(\Theta)$ is constant with respect to the change of variables $\rho'_{sj}(y_j) = \rho_{sj}(y_j) + c$ and similarly for ϕ, so we cannot hope to recover $\Theta^\star$. A popular fix for this issue is to drop the last level of ρ and ϕ, so they are only indicators over $L-1$ levels instead of L levels. This allows for the model to be identifiable, but it results in an asymmetric formulation that treats the last level differently from other levels.

Instead, we will maintain the symmetric formulation by introducing constraints. Consider the problem

$$\begin{aligned} \underset{\Theta}{\text{minimize}} \;\; & \ell(\Theta) + \lambda \sum_{g \in G} \|\Theta_g\|_2 \\ & \text{subject to } C\Theta = 0. \end{aligned} \tag{4.10.9}$$

The matrix C constrains the optimization variables such that

$$\sum_{y_j} \rho_{sj}(y_j) = 0$$
$$\sum_{y_j} \phi_{rj}(y_r, y_j) = 0.$$

The group regularizer implicitly enforces the same set of constraints, so the optimization problems of Equation (4.10.9) and Equation (4.6.1) have the same solutions. For our theoretical results, we will use the constrained formulation of Equation (4.10.9), since it is identifiable.

We first state some definitions and two assumptions from Lee et al. (2013b) that are necessary to present the model selection consistency results. Let A and I represent the active and inactive groups in Θ, so $\Theta_g^\star \neq 0$ for any $g \in A$ and $\Theta_g^\star = 0$ for any $g \in I$. The sets associated with the active and inactive groups are defined as

$$\mathcal{A} = \{\Theta \in \mathbb{R}^d : \max_{g \in G} \|\Theta_g\|_2 \leq 1 \text{ and } \|\Theta_g\|_2 = 0,\, g \in I\}$$
$$\mathcal{I} = \{\Theta \in \mathbb{R}^d : \max_{g \in G} \|\Theta_g\|_2 \leq 1 \text{ and } \|\Theta_g\|_2 = 0,\, g \in A\}.$$

Let $M = span(\mathcal{I})^\perp \cap Null(C)$ and P_M be the orthogonal projector onto the subspace M. The two assumptions are

1. Restricted Strong Convexity. We assume that

$$\sup_{v \in M} \frac{v^T \nabla^2 \ell(\Theta) v}{v^T v} \geq m \tag{4.10.10}$$

for all $\|\Theta - \Theta^\star\|_2 \leq r$. Since $\nabla^2 \ell(\Theta)$ is lipschitz continuous, the existence of a constant

m that satisfies (4.10.10) is implied by the pointwise restricted convexity

$$\sup_{v \in M} \frac{v^T \nabla^2 \ell(\Theta^\star) v}{v^T v} \geq \tilde{m}.$$

For convenience, we will use the former.

2. Irrepresentable condition. There exist $\tau \in (0, 1)$ such that

$$\sup_{z \in \mathcal{A}} V(P_{M^\perp}(\nabla^2 \ell(\Theta^\star) P_M (P_M \nabla^2 \ell(\Theta^\star) P_M)^+ P_M z - z)) < 1 - \tau, \tag{4.10.11}$$

where V is the *infimal convolution* of the gauge ρ_I, $\mathcal{I}$ $\rho_I(x) = \inf\{t : t > 0, tx \in \mathcal{I}\}$, and $\mathbb{1}\left[Null(C)^\perp\right]$:

$$V(z) = \inf_{z = u_1 + u_2} \{\rho_I(u_1) + \mathbb{1}\left[Null(C)^\perp\right](u_2)\}.$$

Restricted strong convexity is a standard assumption that ensures the parameter Θ is uniquely determined by the value of the likelihood function. Without this, there is no hope of accurately estimating $\Theta^\star$. It is only stated over a subspace M which can be much smaller than $\mathbb{R}^d$. The Irrepresentable condition is a more stringent condition. Intuitively, it requires that the active parameter groups not be overly dependent on the inactive parameter groups. Although the exact form of the condition is not enlightening, it is known to be necessary for model selection consistency in lasso-type problems (Zhao and Yu, 2006; Lee et al., 2013b) and a common assumption in other works that establish model selection consistency (Ravikumar et al., 2010; Jalali et al., 2011; Peng et al., 2009). We also define the constants that appear in the theorem:

1. Lipschitz constants L_1 and L_2. Let $\Lambda(\Theta)$ be the log-partition function. $\Lambda(\Theta)$ and $\ell(\Theta)$ are twice continuously differentiable functions, so their gradient and hessian are locally Lipschitz continuous in a ball of radius r around $\Theta^\star$:

$$\|\nabla \Lambda(\Theta_1) - \nabla \Lambda(\Theta_2)\|_2 \leq L_1 \|\Theta_1 - \Theta_2\|_2, \ \Theta_1, \Theta_2 \in B_r(\Theta^\star)$$
$$\left\|\nabla^2 \ell(\Theta_1) - \nabla^2 \ell(\Theta_2)\right\|_2 \leq L_2 \|\Theta_1 - \Theta_2\|_2, \ \Theta_1, \Theta_2 \in B_r(\Theta^\star)$$

2. Let $\overline{\tau}$ satisfy

$$\sup_{z \in \mathcal{A} \cup \mathcal{I}} V(P_{M^\perp}(\nabla^2 \ell(\Theta^\star) P_M (P_M \nabla^2 \ell(\Theta^\star) P_M)^+ P_M z - z)) < \overline{\tau}.$$

V is a continuous function of z, so a finite $\overline{\tau}$ exists.

Theorem 4.10.1. *Suppose we are given samples $x^{(1)}, \ldots, x^{(n)}$ from the mixed model with unknown parameters $\Theta^\star$. If we select*

$$\lambda = \frac{2\sqrt{256 L_1}\overline{\tau}}{\tau} \sqrt{\frac{(\max_{g \in G} |g|) \log |G|}{n}}$$

and the sample size n is larger than

$$\max \begin{cases} \frac{4096 L_1 L_2^2 \overline{\tau}^2}{m^4 \tau^4} \left(2 + \frac{\tau}{\overline{\tau}}\right)^4 (\max_{g \in G} |g|) |A|^2 \log |G| \\ \frac{2048 L_1}{m^2 r^2} (2 + \frac{\tau}{\overline{\tau}})^2 (\max_{g \in G} |g|) |A| \log |G|, \end{cases}$$

then, with probability at least $1 - 2\big(\max_{g \in G} |g|\big) \exp(-c\lambda^2 n)$, the optimal solution to (4.6.1) is unique and model selection consistent,

1. $\|\hat{\Theta} - \Theta^\star\|_2 \leq \frac{4}{m} \left(\frac{\overline{\tau}+1}{2\tau}\right) \sqrt{\frac{256 L_1 |A| (\max_{g \in G} |g|) \log |G|}{n}}$,
2. $\hat{\Theta}_g = 0$, $g \in I$ *and* $\hat{\Theta}_g \neq 0$ *if* $\left\|\Theta^\star_g\right\|_2 > \frac{1}{m} \left(1 + \frac{\tau}{2\overline{\tau}}\right) \sqrt{|A|} \lambda$.

Remark 4.10.2. *The same theorem applies to both the maximum likelihood and pseudolikelihood estimators. For the maximum likelihood, the constants can be tightened; everywhere L_1 appears can be replaced by $L_1/128$ and the theorem remains true. However, the values of $\tau, \overline{\tau}, m, L_1, L_2$ are different for the two methods. For the maximum likelihood, the gradient of the log-partition $\nabla\Lambda(\Theta)$ and hessian of the log-likelihood $\nabla^2 \ell(\Theta)$ do not depend on the samples. Thus the constants $\tau, \overline{\tau}, m, L_1, L_2$ are completely determined by $\Theta^\star$ and the likelihood. For the pseudolikelihood, the values of $\tau, \overline{\tau}, m, L_2$ depend on the samples, and the theorem only applies if the assumptions are made on sample quantities; thus, the theorem is less useful in practice when applied to the pseudolikelihood. This is similar to the situation in Yang et al. (2013), where assumptions are made on sample quantities.*

Bibliography

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2011). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4:1–106.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516.

Beck, A. and Teboulle, M. (2010). Gradient-based algorithms with applications to signal recovery problems. *Convex Optimization in Signal Processing and Communications*, pages 42–88.

Becker, S., Candès, E., and Grant, M. (2011). Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, pages 1–54.

Belloni, A., Chernozhukov, V., et al. (2011). Supplementary material for l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Belloni, A., Chernozhukov, V., et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81.

Benjamini, Y., Yekutieli, D., et al. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41(2):802–837.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The statistician*, pages 179–195.

Bühlmann, P. L. and van de Geer, S. A. (2011). *Statistics for High-dimensional Data*. Springer.

Chen, D. and Plemmons, R. J. (2009). Nonnegativity constraints in numerical analysis. In *Symposium on the Birth of Numerical Analysis*, pages 109–140.

Cheng, J., Levina, E., and Zhu, J. (2013). High-dimensional mixed graphical models. *arXiv preprint arXiv:1304.2810*.

Combettes, P. and Pesquet, J. (2011). Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212.

Edwards, D. (2000). *Introduction to graphical modelling*. Springer.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 74(1):37–65.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Foygel Barber, R. and Candes, E. (2014). Controlling the false discovery rate via knockoffs. *arXiv preprint arXiv:1404.5609*.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2008a). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2008b). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010a). Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Technical Report, Stanford University.

Friedman, J., Hastie, T., and Tibshirani, R. (2010b). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In *Computer Sciences and Statistics Proceedings of the 23d Symposium on the Interface*, pages 571–578. Defense Technical Information Center.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Joint structure estimation for categorical markov networks. *Submitted. Available at http://www. stat. lsa. umich. edu/˜ elevina*.

Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research*, 10:883–906.

Jalali, A., Ravikumar, P., Vasuki, V., Sanghavi, S., ECE, U., and CS, U. (2011). On learning discrete graphical models using group-sparse regularization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*.

Kim, S., Sohn, K.-A., and Xing, E. P. (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):i204–i212.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. The MIT Press.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lauritzen, S. (1996). *Graphical models*, volume 17. Oxford University Press, USA.

Lee, J., Sun, Y., and Saunders, M. (2012). Proximal newton-type methods for minimizing convex objective functions in composite form. *arXiv preprint arXiv:1206.1623*.

Lee, J. D. and Hastie, T. J. (2014). Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, (just-accepted):00–00.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2013a). Exact post-selection inference with the lasso. *arXiv preprint arXiv:1311.6238*.

Lee, J. D., Sun, Y., and Taylor, J. (2013b). On model selection consistency of m-estimators with geometrically decomposable penalties. *arXiv preprint arXiv:1305.7477*.

Lee, J. D. and Taylor, J. E. (2014). Exact post model selection inference for marginal screening. *arXiv preprint arXiv:1402.5596*.

Lee, S., Ganapathi, V., and Koller, D. (2006). Efficient structure learning of markov networks using l1regularization. In *NIPS*.

Liang, P. and Jordan, M. (2008). An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *Proceedings of the 25th international conference on Machine learning*, pages 584–591. ACM.

Liu, Q. and Ihler, A. (2011). Learning scale free networks by reweighted l1 regularization. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Liu, Q. and Ihler, A. (2012). Distributed parameter estimation via pseudo-likelihood. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Lockhart, R., Taylor, J., Tibshirani, R., and Tibshirani, R. (2014). A significance test for the lasso (with discussion). *Annals of Statistics*.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Methodological)*, 72(4):417–473.

Meinshausen, N. et al. (2013). Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics*, 7:1607–1631.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746.

Ravikumar, P., Wainwright, M., and Lafferty, J. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Reid, S., Tibshirani, R., and Friedman, J. (2013). A study of error variance estimation in lasso regression. *Preprint*.

Rodriguez-Yam, G., Davis, R. A., and Scharf, L. L. (2004). Efficient gibbs sampling of truncated multivariate normal with application to constrained linear regression. Technical report, Department of Statistics, Colorado State University.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Schmidt, M. (2010). *Graphical Model Structure Learning with l1-Regularization*. PhD thesis, University of British Columbia.

Schmidt, M., Kim, D., and Sra, S. (2011). Projected newton-type methods in machine learning.

Schmidt, M., Murphy, K., Fung, G., and Rosales, R. (2008). Structure learning in random fields for heart motion abnormality detection. *CVPR. IEEE Computer Society*.

Slawski, M., Hein, M., et al. (2013). Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, 7:3004–3056.

Taylor, J., Lockhart, R., Tibshirani, R. J., and Tibshirani, R. (2014). Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889*.

Taylor, J., Loftus, J., and Tibshirani, R. J. (2013). Tests in adaptive regression via the kac-rice formula. arXiv:1308.3020. Submitted.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490.

Tur, I. and Castelo, R. (2012). Learning mixed graphical models from data with p larger than n. *arXiv preprint arXiv:1202.3765*.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636.

Wright, S., Nowak, R., and Figueiredo, M. (2009). Sparse reconstruction by separable approximation. *Signal Processing, IEEE Transactions on*, 57(7):2479–2493.

Yang, E., Ravikumar, P., Allen, G., and Liu, Z. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems 25*, pages 1367–1375.

Yang, E., Ravikumar, P., Allen, G., and Liu, Z. (2013). On graphical models via univariate exponential family distributions. *arXiv preprint arXiv:1301.4183*.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 76(1):217–242.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67(2):301–320.

Jason Dean Lee

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Trevor J. Hastie) Principal Co-Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Jonathan E. Taylor) Principal Co-Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Lester Mackey)

Approved for the Stanford University Committee on Graduate Studies

---